SCIENTIFIC DECISION MAKING IN BUSINESS

ABE SHUCHMAN

Scientific Decision Making in Business

Readings in Operations Research for Nonmathematicians

ABE SHUCHMAN

Columbia University

Holt, Rinehart and Winston, Inc.

New York Chicago San Francisco Toronto London

To

LOTTIE

Printed in the United States of America
Library of Congress Catalogue Card Number: 63-8145
27885-0113

Preface

Orbiting satellites, atomic reactors, synthetic fibers, and tranquilizers make it readily apparent that we live in an era of very rapid and accelerating technological change. Moreover, most business executives as well as students of business are aware that this technological change has not been confined solely to materials, products, and production processes but has extended to management itself. They sense, in fact, that a transformation in the managerial art is under way, which may have as profound an effect on business as the harnessing of atomic fusion has had on the art of war.

Unfortunately, as I have found time and again, although executives and students, by and large, recognize that a new and more powerful technology of managerial decision making has emerged and is growing rapidly, they have little understanding of either its nature or its promise. This has not been the result, however, of a lack of effort. It has been, instead, the consequence of a lack of the equipment needed to comprehend the available books on the subject.

The new management technology, which has been named operations research or management science, is in a fundamental sense a branch of applied mathematics, and its practitioners have been largely mathematicians, physical scientists, and mathematical economists. As a consequence, the books that have been written with the aim of introducing present and prospective executives to management science have, on the whole, presumed more mathematical sophistication and skill than is possessed by most managers and students in schools of business. On the one hand, therefore, these books, although excellent in many respects, have not become the vehicles for the more widespread understanding that their authors had undoubtedly hoped they might be. And, on the other hand, managers and students have been unable to satisfy fully their desire for greater knowledge.

This gap in the textbook literature of management science has come to my attention forcefully and often in my role as a teacher both in a school of business and in executive development programs. To cope with the situation, I have endeavored, over the past five years, to search closely the periodical literature of the discipline for articles and other items that, although demanding very little mathematical equipment, nevertheless succeed

in conveying the sense and substance of management science and its various techniques. This search, as I hope the readers of this book will agree, has been rewarded. To my good fortune and that of my students and others seeking better acquaintance with management science, many practitioners of the discipline are not only creative researchers but also very talented in communicating the essentials of their knowledge to those who have little mathematical competence.

This volume is a collection of the best writing I have been able to find that describes the aims, methods, and tools of management science or operations research without recourse to technical jargon or complex mathematical symbolism. In addition, I have included a group of selections that describe the application of these methods and tools to specific problems of production, marketing, and finance. Most of the items included have been written by active and outstanding practitioners of the discipline but a few have been drawn from the works of men who are primarily statisticians; and where I have been unable to find a suitable article dealing with an important aspect or tool of the discipline, I have tried to fill the gap myself. Beyond this, I have also sought to supply a framework that may enable the reader to comprehend more readily how the parts of the book fit together and why the articles have been grouped as they have been.

The audience I have had constantly in mind as I have prepared this volume has been, as I have suggested, that consisting of students in schools of business and working executives who have little knowledge of mathematics and statistics but want somewhat more than an extremely rudimentary understanding of the what, why, and how of management science. It is my hope, in fact, that this volume will be found useful as a textbook for undergraduate and graduate courses in operations research or management science given in schools of business and for executive development programs. In addition, it may serve as a convenient source of background reading for any other course in a school of business into which it is desired to introduce relevant applications of operations research. Finally, it is my hope that many executives will find the book can take an important place in their programs for personal development.

Being a collection, this volume is obviously the work of many people. I am deeply indebted to them all and wish to express my thanks to the authors and publishers who so kindly granted me permission to reproduce the selections it contains. Without their cooperation, this book would not have been possible.

I am grateful also to my wife for her encouragement, assistance, and advice during the preparation of the book.

New York
May, 1963

Abe Shuchman

Contents

Introduction

This book is about a war baby. It concerns a method for attacking and solving complex problems which was fathered by military necessity and is being reared to maturity by business as well as military needs. It is about an applied science known as Operations Research, which may well become one of the truly powerful tools for unraveling intricate business problems.

To most business executives, Operations Research is still little more than a name. There is a growing belief, however, that this situation cannot continue. Operations Research is concerned with the very core of a manager's job—decision making. It is a method and set of techniques derived from the physical sciences and mathematics, which promises to effect considerable improvement in the quality of managerial decisions. Although still a very young discipline, it has already provided insights into complicated business problems such as inventory control, production scheduling, warehousing, and advertising which executives did not previously have. It has already made possible better and even "best" solutions for many such problems. Executives, who must be concerned with the quality of their decisions, cannot, therefore, remain content with the largely incidental and fragmentary information about Operations Research which they now possess.

The survival of a firm in a competitive business world depends above all upon the decisions which its executives make. These decisions determine the types and quantities of resources which a firm is to have and the uses of these resources. These are life and death questions for a firm. If the executives' decisions are "right" a firm will prosper and grow. If, however, they are inferior to those of competitors not only will the firm's market position deteriorate but the firm may very well die. In the business world, therefore, the value of an executive is determined largely by the quality of his decisions. In self defense, if for no other reason, executives must become familiar with any group of ideas that may help them make better decisions more frequently.

Operations Research is such a group of ideas. Executives, prospective and present, in training and on the firing line, must make the effort to understand the discipline. This understanding cannot and need not be that of an expert, but it must be sufficient to enable the executive to know when and how the expert can assist him. And it must be deep enough so that he can help the expert to help him.

The understanding of Operations Research needed by executives is, it seems, a basic understanding of its central ideas, language, tools, accomplishments, promise and limitations. And this book seeks to provide it by giving ready access to nontechnical writings of experts in the field which have appeared in widely scattered sources. Understanding may be facilitated also by knowledge of the origins of Operations Research and of its relationship to executive decision making. Consequently, before allowing the experts to speak, it may be profitable to examine briefly the history of Operations Research and the grounds for asserting that it can improve the quality of executive decisions.

THE ORIGINS OF OPERATIONS RESEARCH

In a broad sense, Operations Research is not a radical innovation. It is, instead, another phase during this century in the development of the stream of ideas associated with "scientific management." Moreover, many of the techniques now identified with Operations Research have existed quite independently for some time. In a narrower sense, however, Operations Research is a new discipline. Its method, concepts and techniques, although derived from the physical sciences and mathematics, had almost never before World War II been applied in an organized and systematic manner to the operating problems of human organizations, military, political, or industrial. Viewed from this perspective, therefore, Operations Research can be regarded as a war baby.

The first use of organized Operations Research appears to have occurred during the "Battle of Britain" in 1941. British scientists had just developed radar, but the military, unfamiliar with this novel detection device, were uncertain about how it might best be used. They, therefore, enlisted the aid of the men who had designed the equipment. These men, as scientists, had been trained and were highly skilled in attacking problems in a definite manner. They approached the radar application problem in the same way. They carefully collected operating data and analyzed these with refined and powerful mathematical and numerical techniques. Using the results of their analyses, they developed a theory to explain the data which had been collected. Then they used the facts and theory to make predictions about future operations. And finally, they tested their predictions against data from further operations and modified their theory so that future predictions might correspond more closely to actual results. In short, the radar problem was studied using the scientific method as it is used in the natural sci-

ences; in other words, in conjunction with advanced techniques of quantitive analysis.

As a result of the work of these scientists, it has been estimated that British power in air defense during the "Battle of Britain" increased tenfold. This was so astounding a success that the military proceeded to organize additional teams of scientists to study other problems of weapon use, tactics, and strategy. One such team was assigned the study of the use of depth charges by naval forces in anti-submarine attacks. This team found that the charges were always set to detonate at a depth of 100 feet because, at this depth, the pressure of the water caused the detonation to have its greatest power. The correctness of this setting was, however, doubted. They noted that although the setting resulted in greatest detonating power it also delayed the launching of an attack, since it was necessary to wait until the diving submarine reached a depth of 100 feet. During this delay, the submarine might turn in any direction unknown to the attacker, and might thus be far from the point of detonation.

In order to come to grips with this problem, the scientists employed probability theory to calculate the chances that the submarine would be within lethal range of the detonation for various settings of the depth charge. In each case they assumed that after submerging, the submarine might follow any course vertically and horizontally. These calculations made it clear that the traditional setting of 100 feet was too deep and that, in theory, a reduction of the setting to 35 feet would greatly increase the chances of a "kill." When the scientists' theory was finally tested, after much resistance and discussion, it was wonderfully validated in practice. The number of sinkings reported increased almost exactly in proportion to their predictions.

One last example may further illuminate the nature of the work done and the contributions made by British Operations Research scientists during World War II. Early in the war the British air force used ordinary bombs against enemy submarines. These were effective only if a direct hit occurred because they exploded on the surface of the water and could not penetrate a submarine's pressure hull unless they struck the submarine's deck. To obtain underwater explosions more destructive to submarine hulls, depth charges were adapted for use by airplanes. However, there was disagreement about the depth at which the charges should be set to explode. Some squadrons used a setting of 150 feet, apparently because they felt that an attacked submarine was most likely to be submerged. On the other hand, other squadrons used a setting of 50 feet. Since submarines at a depth of 150 feet could not be seen and so could not be attacked, while submarines near the surface which could be seen and attacked would hardly be damaged by an explosion at the 150 foot depth, the absurdity of the deep setting was manifest and the shallow setting became doctrine. However, the argument between the "deep-setters" and the "shallow-setters" continued, and finally, a team of scientists was called upon to study the problem.

The scientists soon concluded that the critical variable was the position of the

submarine at the moment the depth charge was dropped. If most attacks were made on surfaced submarines, then even a setting of 50 feet was too deep, since the explosion would occur too far away to have a chance of producing lethal damage. But if most attacks were made on submarines which had just dived, then the 50 foot setting might be satisfactory. Whether or not it was satisfactory depended upon the accuracy that could be achieved in bombing submerged submarines.

To determine this accuracy, the scientists compiled and analyzed operational data with the analysis consisting of the calculation of the probability of a successful attack when the submarine was submerged to various depths. This probability they found to be very, very low, regardless of the setting used. Therefore, they urged that if the probability of a successful attack was to be maximized, the setting used would have to be that which was best for attacks on surfaced submarines, unless this type of attack hardly occurred.

Thus the key question was the relative frequency of attacks on surfaced submarines. Analysis of the operational data revealed that 40 per cent of all attacks were made on fully surfaced submarines and another 10 per cent were made when a part of the submarine was visible. It was clear that the 50 foot setting was unsatisfactory. And in fact, the probability calculations indicated that a reduction in the setting of the depth charge to 25 feet would at least triple the chance of success in the average attack. This change was recommended along with a rule that no charge was to be dropped if the submarine had already been submerged for more than half a minute. The recommendations were adopted and in a short time the effectiveness of airplane anti-submarine attacks more than doubled.

The achievements just enumerated as well as many more did not go unnoticed on this side of the Atlantic. Consequently, when the United States entered the War in December 1941, our military quickly emulated the British and organized Operations Research teams. These American teams "tackled" and successfully solved a great variety of problems, many of which had been believed to be intractable to the methods and high powered mathematical techniques of scientific research. And success led to a steady increase in the use of such teams throughout the war. Among the many contributions made were improved methods of search for submarines, better arrangements for convoys and their escorts, methods for achieving better aerial bombing accuracy, and appropriate tactics or maneuvers in the face of Kamikaze or suicide plane attacks. The enduring effect of the work of these Operations Research teams is that the military, which was at first highly sceptical, now accepts the usefulness and value of Operations Research without reservation. Operations Research has become a permanent and important part of our military apparatus.

After the war, many of the scientists who had participated in military Operations Research returned to peace time activities with the conviction that the method

and techniques which they had used to improve military operations could be used with equal success to improve business operations. Business did not meet them with open arms. The business boom which followed the war was based on a pent-up consumer demand for goods which seemed insatiable. Executives emphasized production at any cost rather than efficiency. Better or best solutions to problems were of little concern as long as there was at hand a solution, any solution that permitted continued output. Even after production capabilities caught up with and then exceeded demand, even after executives had once again to stress efficiency, business did not rush into adopting Operations Research. A more fundamental barrier than market conditions existed and is still formidable today. This is the lack of understanding of the nature, methods, and aims of Operations Research. This lack of understanding and even misunderstanding, coupled with a more or less inherent distrust of "long-haired intellectuals" engendered considerable suspicion and scepticism among executives. On the one hand, they said, "I can't use it because my business is different; everything is so uncertain," or "I can't use it because there are too many intangibles like good will, tastes and human relations in my business." On the other hand, however, as some executives tried Operations Research and as evidence accumulated showing it could be used with profit in highly complex business situations involving intangibles, many executives retreated to, "It's too expensive; only the very largest, 'best heeled' companies can afford it."

Despite this general resistance born of apathy and lack of understanding, there has been in recent years a rapid growth in the number of firms selling Operations Research as a service to industry as well as in the number of firms which have formed their own Operations Research teams. Competition is a severe teacher. It is not concerned with desires and attitudes but with results. And when firms have reduced costs and increased sales by using information provided by operations researchers, the executives of competing firms have been compelled to reconsider the value of the discipline. At the same time, the considerable efforts of professional groups such as the Operations Research Society of America, the Institute of Management Sciences, the Society for Advancement of Management, and the American Management Association, the development of special courses at universities such as the Case Institute of Technology and the Massachusetts Institute of Technology, and the activities of consulting firms have accomplished much by way of spreading information about Operations Research throughout industry. Gradually, executives in increasing numbers are striving for and gaining an understanding of Operations Research. Today, an estimated 5 per cent of the nation's business organizations either have their own Operations Research department, employ outside Operations Research consultants, or use a combination of both. Even more significantly, 60 of the nation's 100 largest firms, including such industrial and commercial leaders as General Motors, General Electric, Standard Oil of New Jersey, United Airlines, Sears Roebuck and the Bank of

America, now have Operations Research departments. The discipline is slowly coming of age as a management tool and as it matures it will undoubtedly come into ever wider use in American business.

OPERATIONS RESEARCH AND EXECUTIVE DECISIONS

The large salaries commanded by business executives who can consistently make decisions of high quality testifies in our economy to the scarcity of such talent. It suggests also that making decisions of high quality consistently is a difficult job. True appreciation of the difficulty of the job requires understanding of the nature of decision making.

In business, as in human activity, generally, there is usually more than one way to "skin a cat." Rarely does a business problem permit one and only one solution. Ordinarly, there are many possible solutions. Not all solutions, however, are equally good. In making a decision an executive must, therefore, do two things. First, he must define each of the possible courses of action which appear to be feasible solutions for his problem. Then, he must choose from among these the one which is the "best" solution. This need for choosing, for selecting one course of action from among a number of possible alternatives is the most distinctive characteristic of decision making.

Decision making in business usually has two other important characteristics. It requires the executive to project himself into the future and it also involves uncertainty. To see this, consider an executive who wants to choose the "best" from among many possible solutions to a problem. How must he proceed? First, if he is to choose the "best" solution, he must, obviously, be able to recognize it. He must define "best." Such a solution is always one which ensures fullest achievement of the executive's objectives. In defining "best," therefore, the executive must really define the goals which he wishes to achieve through solving his problem. Having done this, the executive must then predict the outcome of each of his alternative courses of action in terms of these goals. Then, he compares these outcomes. And finally, after he has made the comparisons, he is ready to choose the "best" solution.

Thus, the selection of a "best" solution requires that the executive "know" the outcome of each of his possible courses of action. This outcome depends on circumstances not as they were, or are, but as they will be when the course of action is implemented. It depends on the events of tomorrow rather than of yesterday and today. In making his choice, the executive is concerned, therefore, with the future. And it is this orientation toward the future which is a second characteristic of decision making in business.

An executive cannot know precisely what will happen tomorrow. He rarely has full and perfect information about the past or present and never about the future. His predictions about the outcome of each of the alternatives open to him

must be conjectural. They must contain an element of uncertainty. Decision making in business, consequently, involves "taking a chance." It involves risk.

The principal job of an executive is then to make choices between alternative courses of action by reading a future about which he is usually uncertain. No wonder that making decisions of high quality consistently is difficult! Experience, rules of thumb and intuition must be called upon. The executive must make guesses and "play" hunches. Guesses may be "wild" or informed. They may constitute the entire basis for a decision or they may only fill in where information is inadequate or unavailable. The real question an executive must answer is not whether he is willing to base his decisions on guesses or hunch, for he cannot avoid this entirely. It is, instead, the question of the extent to which he is willing to do so.

Increasingly, executives have concluded that extensive reliance on vague rules of the trade, "feel of the situation" and intuition is too dangerous. The growth in the size and complexity of enterprise and the accelerated change of pace in the environment have made the executive's problems vastly more complicated. The organization and operations of many firms have become so complex, in fact, that only rarely can their executives see directly to the heart of a problem. More commonly, they cannot even readily define a problem, much less the many critical factors involved, the inter-relationships between these factors, the possible courses of action and their outcomes and the probability of occurrence of each outcome. Under these circumstances, decision making "by guess and by gosh" has become extraordinarily risky. Too much is lost and too many people are injured if a firm's executives make decisions inferior to those of competitors. And the executives of competing firms may very well make better guesses or have better hunches. Or, and this is a much more compelling consideration, they may even have found a method which enables them to make better decisions consistently.

In this situation, many executives have endeavored to reduce their reliance on intuitive vision in making decisions. They have tried to substitute a systematic and rational approach to problem solving. These executives seek, consciously and explicitly, to define a problem and the factors relevant to its solution. They specify the objectives and conditions which a solution must satisfy. They make an effort to mobilize information about resources and the environment and to define possible solutions. They try to estimate and compare the costs, benefits and risks of each possible solution. And finally, they select the solution which they regard as the best balance of cost, benefit and risk. Throughout this process, inspiration and judgment remain important, but the aim is to make them the plus ingredient in decision making rather than its entire foundation.

To some extent, this systematic and rational procedure has been rewarding. Managerial decisions have often been improved. The decisions which have resulted have generally been superior to those which have come "off the top of the head." The procedure has not been, however, the spectacular success that had

been hoped. Decision making has not become much easier for the executive. It has often, in fact, become more difficult. The improvement in the quality of decisions has made "get the facts" a byword in executive chambers and many firms have, consequently, become voracious collectors and processors of information. The volume and diversity of information moving across the desks of their executives has attained titanic proportions. The effect has frequently been, however, that the executives' problems have been obscured rather than illuminated. Effects have been reported but not their causes. Symptoms have been described but not their etiology. Rapid and careful sifting of the information to cull out the significant and relevant has become enormously time consuming. The precise definition of a problem and the methodical evaluation of alternative solutions, in a complex situation, has come to require more time and greater capabilities than the executives had to invest. And the uncertainties and risks involved in alternative solutions have remained, despite the flood of data, exceedingly crude estimates based on opinion and vague generalities. The upshot of this has been that executives must still "fly blind" and rely upon experience and intuition in most important situations. Although the systematic and rational approach to problem solving has been of some help in most situations and even of great help in some situations, experience and inspiration have continued to be the principal sources of decisions.

The advent of Operations Research in business promises to alter this situation. The theory building and testing approach to problem solving based on advanced techniques of quantitative analysis greatly enhances the ability to isolate critical variables in a situation and to relate events to them in a simple and cogent manner. It often makes the inference possible from experience and the torrent of available data of meanings and relationships that are not at all apparent or common sense. Again, in disputes concerning which of two alternative qualitative views is better, this approach can often resolve the dispute in a statement such as, "Action A is x per cent better than Action B for New York and y per cent worse for Los Angeles." Finally, even when the approach results in the conclusion that the common sense view held by executives is correct, as is often likely to be the case, it will usually be able to buttress the view with numerical proof and so give added confidence to the decisions based on it.

To be more specific, the method and techniques of Operations Research can ease the executive's difficulties in making decisions by contributing the following:

1. A better and more logical description of his objectives and of the assumptions on which they are based;
2. A more precise and illuminating definition of his problem and of the critical factors involved, the relative importance of each, and the relationships among them;
3. A clear indication of the information required in order to determine the "best" solution;
4. The ability to take into account, in determining the "best" solution, a larger number of relevant factors;

5. A precise description of many more of the possible solutions for the problem, the assumptions underlying each, and the costs, benefits, and risks involved in each;
6. The ability to compare many more possible solutions and to locate the "best" among them, rapidly, efficiently, and with considerable confidence;
7. A basis for predicting the consequences of changes in his firm's procedures or in the environment.

In summation, Operations Research promises to provide executives with a more precise description of the assumptions, cause-and-effect relationships and risks at the root of business operations. It promises to convert many problems which now seem too complex, too chaotic, too random, or too uncertain for treatment with anything other than intuition, experience and judgment to ordered patterns which can be analyzed with a vast variety of tools already known and widely used in other disciplines. It promises, thus, to provide an understanding of the underlying characteristics and fundamental nature of many knotty problems, an understanding which should give managers new insights, and the capability to determine better, if not the best, solutions with greater speed and assurance. What Operations Research promises is, to put it squarely, a considerable improvement in the quality of managerial decisions.

If this promise is realized in any considerable measure—and accumulating evidence suggests that it will be—then Operations Research will undoubtedly replace intuition and experience as the cornerstone of decision making in business. As executives find that many problems which had been intractable become more manageable after analysis with the methods and techniques of Operations Research, they will surely make a further shift of emphasis in decision making from hunch and inspiration to the more systematic and rational procedure. Accordingly, the advent of Operations Research may very well mean that managerial decision making is to be, in the future, less of an art and more of a science. And if this is the possible significance of Operations Research, then students of business, both in school and on the firing line, must examine it carefully.

part ✦ 1

What Is Operations Research?

It is often helpful on approaching a new subject to start with a survey which focuses on broad outlines rather than details. This survey seeks to answer general questions about the subject such as: What is it? What does it do? How does it do this? Why does it do this in this way? It seeks to give the student a basic familiarity with the aims of the discipline, the assumptions on which it is based, the kinds of problems with which it is concerned, the approach to these problems which it prescribes and the principal concepts and techniques which it utilizes. It provides, thus, a frame of reference that should enable the student, as he moves to closer examination of the subject, to understand more readily where each detail fits and why. The result is frequently a keener appreciation of the value and limitations of the methods and techniques of the discipline.

The six selections in Part One of this book constitute such a survey. They are, in effect, a "once over lightly" of OR. The first selection was written by the president of The Ramo-Wooldridge Corporation who is also an eminent research scientist. In the article, the author discusses the new relationship between the scientist and the business executive which is implied by the growing practice of OR. In addition, he traces the reasons behind the changed attitude of scientists toward research in military and business problems. And finally, although recognizing the reservations managers have about the utility of scientists in solving operational problems, he points out the potential benefits that could accrue from the acceptance of them in this role.

The second selection was authored by two operations researchers at Arthur D. Little, Inc., the first management consulting firm to offer OR as a service to business. The selection defines OR and some of its fundamental concepts, distin-

guishes OR from other tools of management, and examines some of the contributions and limitations of the discipline.

The next three selections are excerpts from brochures published by the General Electric Company. Their author is senior consultant in the corporation's Operations Research Consulting Service. The first two of these selections discuss the trends and intellectual currents in business which have led to the emergence of OR as a management tool. In addition, they describe the basic aims of OR and the general nature of the managerial problems which OR can help solve. The third of this group of selections is, on the other hand, a description of the basic elements of the procedure involved in an OR attack on and solution of a problem.

The final selection in this segment of the book is a portion of an article written by the Director of the OR Group at the Case Institute of Technology, a pioneer in the development of formal courses of study in OR. This excerpt provides another view of the major phases of an OR project.

✦✦✦✦✦✦✦✦✦✦✦✦✦ OPERATIONS RESEARCH: *the scientists' invasion of the business world*

DEAN E. WOOLDRIDGE

Anyone who today addresses himself to the subject of operations research feels an almost irresistible urge to start by defining the subject. This is probably because, while he may know what he means by the term, he is not quite sure that the words mean the same thing to everyone else. The need here is to start with some kind of definition of operations research that is suitably tailored to the particular set of prejudices to be promulgated.

AN EXAMPLE

Rather than start with a formal and presumably uninteresting, suitably slanted definition, an episode of the last war will serve to illustrate good operations research. The episode in question concerns a squadron of bombing planes that was stationed on one of the small islands in the Pacific in 1944. These planes were owned and operated by the Air Force, and the story revolves about some difficulties with the United States Navy. It seems that every once in a while one of the Air Force planes would be shot at, and sometimes hit, by naval vessels close to which the planes had occasion to fly. This might have been considered to be just a normal wartime occurrence if it were not for the fact that the incidence of such unfortunate events was considerably

Reprinted from the September–October 1956 issue of The Journal of Industrial Engineering, *7:5, 230–235, Official Publication of the American Institute of Industrial Engineers, Inc., 345 East Forty–Seventh Street, New York 17, New York.*

higher for the bombing squadron in question than it was for other squadrons operating in the same general area.

Finally, after some exchange of correspondence back and forth between the Pacific island and Washington, the problem was narrowed down to the low level of reliability of the IFF equipment in the aircraft. To those who are not familiar with this term, IFF is an abbreviation meaning "Interrogate—Friend or Foe." The term refers to a black box in the aircraft that constituted, in effect, a special kind of radio receiver-transmitter combination. When any aircraft was sighted by a Navy ship, the ship would send out a special type of radio signal. If the aircraft in question was equipped with the allied IFF apparatus, then another specially coded signal would be returned to the interrogating ship, indicating that the aircraft in question was friendly; consequently, guns would not be fired. But for some reason, the IFF equipment was not working very frequently for this particular squadron in the Pacific. The routine efforts of maintenance on the Pacific island and the encouragement, and perhaps threats, sent out from Washington didn't improve the situation. Finally an operations research expert was sent to the Pacific island to perform careful detailed analyses, to localize the trouble, and, hopefully, to fix it.

Within about two weeks of the arrival of the operations research expert on the Pacific island, the encounters between the aircraft of that base and Navy ships dropped in a rapid and spectacular fashion, and furthermore, the incidence of such encounters stayed practically at zero for week after week. Finally it was clear that the problem had been completely solved, and word was sent to the operations research expert to return to Washington. Upon his arrival, he was of course warmly greeted by his associates, who gathered around and congratulated him, and were most anxious to learn what unusual feats of analysis, higher mathematics, and electronics design he had been able to accomplish in such a short period of time to produce such a miraculous result. Since operations research men are always modest, objective scientists, the returning hero was not at all hesitant in explaining clearly what he had done. He said, "Well, after I had been there for a week asking questions and analyzing the subject, tracing electronic circuits, and making statistical calculations, I finally discovered what I thought was the problem. To cure it, I stationed a man at the end of the runway each time the aircraft took off, and as each airplane passed, this man held up a big sign that we had painted. This sign carried on it the words, 'Turn on your IFF equipment.' From that time on we had no more difficulty."

The point in telling this story is to direct our attention away from some of the more sophisticated techniques that are occasionally employed by skilled operations research people to arrive at answers to some of the problems they encounter—the use of mathematical models, the Monte Carlo method, the theory of waiting, linear programming, and the like, and to concentrate our attention upon the broad concept of operations research and what it is trying to do for its customer.

MAXIMIZE EFFECTIVENESS

The legitimate subject matter for operations research includes almost all aspects of any complex organization of men and/or equipment. In the example of the story, the operations research scientist was a trouble shooter. In some way a section of the complex organization of the Pacific Air Force was not operating as effectively as was believed reasonable in terms of the over-all objectives of the mission. The basic assignment of the operations research scientist was to find out what was wrong and to make recommendations, whatever they might be, that might somehow improve the effectiveness of the operation. In the example in question, the highest mathematics required for proper solution of the problem was probably arithmetic, and very likely not much of that was needed, although it was likely that the man who was sent out to the Pacific island was fully equipped to deal competently with difficult statistical problems if such had turned out to be the principal points at issue. Conversely, an even less technical solution than the one told in the story might have been the result of the investigation. For example, the answer could have been that it was necessary to fire the Commander of the group. The point of all this is that the subject matter of operations research is very broad—it is the effectiveness of the complex organization of men and machinery in achieving its over-all objectives, whatever they might be.

If you accept the rough but very broad definition of operations research just developed, then you have a right to ask, "What is the difference between the task of the operations research expert and that of the manager of the enterprise? Isn't the manager generally earning his salary on the basis of conducting the operation of the enterprise in such a way as to make it most effective in terms of the over-all company objectives, and isn't it his job to run down inefficiencies and modify operations and procedures as required in order to maximize effectiveness? What is the operations researcher up to when he comes to the manager of a company and asks for an assignment? Is he simply saying to the manager, in effect, 'Anything you can do, I can do better'?"

TO HELP MANAGEMENT

These are questions to which those of us who are engaged in the selling of operations research have a very high sensitivity. When such questions are asked, we frequently fall all over ourselves to try to reassure the manager that the last thing in the world that we want to do is to give him the impression that we are criticizing him of inefficient management, or to cause him to feel that we are trying to step in and take over from him the job of managing his company. We then generally go into quite an exposition to show what the differences are between operations research and management in an attempt to prove that what we can do is to supplement the manager, but under no circumstances are we really setting out to do the same thing. We frequently defeat our own purposes with this kind

of reaction. A more accurate and much simpler reply to the manager who asks if the operations research people are not representing themselves as being able to do a better job than the manager in some of the functions he is being paid for is just a plain, unembellished "Yes."

Of course, there are some management decision areas that are not very well suited to an operations research approach. It is not likely, for example, that the objective, presumed scientific analysis performed by an operations research team could ever develop convincing arguments to prove to the management of the Salvation Army that they should give up their charitable objectives and go instead into the manufacture of poison gas. In other words, there are certain rules of the game that constitute the reason for existence of the company and its basic aims and ambitions that are not properly the subject matter for objective analysis. Admitting that type of exception, however, it is almost true to say that everything else the manager does in operating a complex organization of people and equipment, and every other kind of decision he makes, are basically suitable subjects for operations research.

Therefore, those of us who are trying to sell operations research should face up to the fact that when we go to a manager with our sales pitch, we are in effect saying to him, "Look, we don't believe that you are managing your company as effectively as it could be managed in terms of your over-all company objectives. Put us on your payroll for a while, and we believe that we can effect an improvement."

REACTIONS

There are undoubtedly operations research people who will not go along with this rather blunt interpretation of what it is that operations researchers are trying to sell to business people, but there is little question as to how most managers feel after listening to the propaganda. By and large, if they don't end up being hopelessly confused about what operations research is all about, they generally arrive at the conclusion that what the operations researcher is trying to do is to tell the manager that he isn't doing the best possible job of management and that operations research can improve his performance. This poses the manager with a very interesting problem. The only thing he can be absolutely sure of, if he authorizes operations research activities either by groups within his company or by outside consultants, is that he will end by spending money for their salaries. No one, least of all the legitimate operations research scientist, can or will guarantee in advance that the results will be worth the expenditure.

There are two ways that managers react to this proposition. One type of reaction is simple, straightforward, and to the point. The manager says either to himself or out loud, with or without profanity, something about as follows. "I'm having a hard enough time keeping this company in the black as it is. The last thing in the world I need is to pay some long-haired scientists to come down out of their ivory towers and tell me how to run my business." This point of view no doubt has some merit. It nicely eliminates one decision that the

manager would otherwise have to make. It prevents a certain amount of confusion in the organization that always results when operations research teams go to work, and it certainly saves the money that would have to be paid for the salaries of the operations research group. It has only one important drawback. If the operations research people were right, the company is not going to get the benefit of improved operations.

Fortunately, a great many managers react in an entirely different way to the suggestion that an operations research team can improve the management of their enterprises. Each of this class of managers goes through something like the following process of reasoning. He starts with the observation that he is sitting on top of quite a complex organization involving hundreds or thousands of people. These people are divided into various groups, each group living in accordance with a certain set of rules, regulations, and procedures characteristic of the group and related to the other groups by another set of rules, regulations, and procedures. There are accountants, production control people, manufacturing people, engineers, purchasing people, budgeting groups, and so on. Usually the manager of the enterprise is thoroughly familiar with what goes on in only one or two of these major sub-empires, and he is so busy that he has difficulty in maintaining himself current even with respect to these one or two areas. In all other major areas of activity, the manager knows that he himself does not well understand everything that goes on. From time to time major problems have arisen in the operation of the company that appeared to be localized in one or another of the functional areas. On some of these occasions, the manager has attempted to probe into the activities of one or more of these groups and learn enough about their detailed operations to permit him to form his own personal conclusions as to what his problems were. Usually he has been frustrated by such efforts, primarily because he has encountered such ramifications, complexities, and multiple sets of human and functional interrelationships that he simply could not accomplish his objective of self-education in the limited time he could spend on the matter. But this kind of manager has the recollection of such events in his mind, and he therefore suspects strongly that there are areas in his company where the methods and procedures governing the day-to-day operational and decision-making activities are not well tailored to the special requirements of his company's business.

This kind of manager is by no means rare. He is typical of the managers of nearly all large and many medium-sized enterprises. Such managers feel frustrated about some aspects of their company operations that they are sure are not as efficient as they should be, but they don't know where to turn to get an impartial, objective appraisal that takes account of broad company aims and properly subordinates the local interest, prejudices, and established traditions that may currently govern some of the company activities.

Such a manager does not feel insulted when it is suggested that some of the operations of his company may be less

efficient than they should be. Furthermore, since he is still convinced that if he personally could only spare the time to dig into the operations he would be able to turn up with major improvements, it is not hard for him to accept the basic idea that an operations research team might be able to accomplish much the same result. Then, too, he knows that it isn't necessary that the operations research team be composed of the world's greatest geniuses, for he is convinced from his own experience that the application of common sense and reason, combined with a proper appreciation of basic company objectives, can go a long way in generating efficient operational and decision-making methods and procedures.

To sell this type of manager on an operations research program, it is mainly necessary to convince him that the team that can be put on the problem will be composed of competent men of reasonable intelligence who can bring to bear an objective, quantitative approach, and that the team, furthermore, will do its work in a sufficiently smooth and tactful manner so as not to produce too much disruption in the day-to-day affairs of the company. If these conditions can be met and if the price isn't too high, then this type of manager is going to buy operations research.

THE SCIENTIST

Our thesis is that almost any complex business involving lots of people is likely to include procedures, methods and ways of making decisions that are not tailored to the actual over-all company objectives as well as they might be. Any team of reasonably intelligent and experienced men, with an objective and quantitative approach to problems can, if given the time and the opportunity, dig into a situation of this sort and turn up with recommendations for improvements that probably will more than pay for the salaries of the operations research team, and occasionally may even result in quite a major increase in the effectiveness of the company.

Operations research has been presented in this way in the hope that it will help clear up a bit of the confusion that sometimes seems to surround this subject. At the same time, however, there may be a little more uncertainty as to the difference between operations research and old-fashioned management consulting. After all, haven't the management consultant firms for years engaged in doing exactly the kind of thing that has been described—going into business and industrial establishments, analyzing their operations, and making recommendations for changes to improve over-all effectiveness?

Here again, to be a bit unorthodox from the point of view of the operations research fraternity, management consulting and operations research are terms that have nearly the same meaning. It is well to add, however, that there are some important differences in the work that has been done in the past under the label of management consulting and the work that is going on today more generally under the label of operations research. The essence of the difference has to do with the professional background of the people who are doing management consulting un-

der the name of operations research. The new groups consist largely of people who have trained as scientists. Probably the most important single fact we must know to understand why we are beginning to hear so much about operations research these days and why it is becoming such a powerful new tool of management is that within recent years scientists have decided that the operations of complex aggregations of men and machines, as typified by business and industrial establishments, constitute legitimate subject matter upon which self-respecting scientists can spend their time.

Those who do not have a technical background may find it difficult to appreciate the significance of this remark. This point is really quite vital to an understanding of what is going on in operations research today. The subject matter we must deal with in trying to understand this situation is essentially a certain kind of snobbishness that characterized scientists. If this sounds like harsh criticism, at least I can say that it is self-criticism as much as anything else, for I started my professional career as a Ph.D. in Physics about twenty years ago, and I believe my attitudes were quite typical. When I came out of graduate school clutching my bright and shining Ph.D. degree, I had had ingrained into my thinking the certainty that there were only a few limited fields that were worthy areas of occupation for a Ph.D. in Physics. As a matter of fact, there was some serious question in those days as to whether there was any field that it was really respectable for a physicist to work in except nuclear physics. In those days nuclear physics did not mean atom bombs, of course. It had to do with learning more about the fascinating and rather mysterious laws of nature that governed the construction and behaviour of the nuclei of atoms.

I remember I had to make a difficult and almost degrading personal decision to accept a position with Bell Telephone Laboratories, where, instead of working on nuclear physics, I found myself doing research on various solid state matters—how electrons behave inside of crystals, and the like. This field of research involved quantum mechanical considerations and in other ways was adequately respectable so that I felt no danger of being shunned by my professional contemporaries for accepting such an assignment. However, it was about as far as a self-respecting physicist could go in those days.

REVELATIONS

Then came the war. The war had a profound psychological effect upon scientists such as myself, for many of us were plunged into problems of a type that we would never in peacetime have considered to be suitable subjects for us to work on. The results were most interesting. To begin with, because we had a good training in basic fundamentals, we turned out to be fairly effective in such matters as the development of radar, electronic computers, and the like, where the necessity of devising entirely new techniques frequently put technical success completely out of the reach of conventional engineering types of approach. The aspect of the situation that was less predictable and therefore

more surprising, however, was that by and large the scientists discovered that this work was interesting and challenging. Pretty soon we learned that the job of inventing radar and computing equipment contained problems that were every bit as difficult and as fascinating to solve as the problems associated with atomic nuclei or solid state physics. This, let me emphasize, was quite a revelation to many of us.

As the war went on, this broadening process continued. Pretty soon, we found ourselves involved not only in the design of new equipment, but also in the problems associated with the use of that equipment in military establishments. We had to concern ourselves with the necessity of providing output data from the equipment that human beings could understand and deal with, perhaps as they were flying fast aircraft. We had to concern ourselves with the problems of maintenance and reliability and with complicated practical problems of logistics and supply. We found ourselves also involved in sales activities, whereby we had to present to basically non-technical people what it was we were trying to do and why we should be permitted to do it. And so it was that literally thousands of men trained as scientists or research engineers, who normally would never have considered working in other than a few highly technical areas, were forced by the exigencies of war to see some of the outside world and to learn that they could find interesting and challenging problems in broad operational situations in which the purely technical content might be fairly small. Now the individuals who went through that experience have, as a consequence, been available for a broad range of assignments since the war.

But something perhaps even more important than that has occurred. Because of the influence that these men have had on the curricula and on the students of the universities where some of them have gone back to work and to teach, and the influence they have exerted through professional societies and the many other ways by means of which scientists keep in touch with one another, the general attitude toward life with which new scientists start out into the world is now quite different than it was before the war.

The effect of this whole process has been quite revolutionary. One of the important practical results is that since the war major weapons systems development projects have come along rapidly because good scientifically trained people have been willing to work on these programs. The consequence of greatest significance for this discussion, however, is that today a man trained as a scientist can still hold up his head professionally and not be looked down upon by his contemporaries if he chooses to go into operations research and concentrate his scientific training on the analysis and solution of problems that arise in large human organizations. Such a situation, unthinkable a few years ago, is a direct consequence of the processes set in motion during World War II.

What is new about management consulting or operations research, then, is that scientists have begun to get into it. It just happens that scientists like the name operations research better than

they do management consulting. (We scientists have not entirely lost our snobbishness, even yet.)

Of what significance to the management of companies is this entry of the scientist into his domain? Is there something here that really presages a higher caliber of management consulting than has been available in the past? Or is all of this talk about operations research simply a consequence of the scientist trying to pat himself on the back? Perhaps there is some of the latter in the picture, but there is also good solid reason for management to be happy about the increasing availability of scientists for operations research assignments.

OBJECTIVITY

The scientifically trained man brings to business management several very important qualities. The first of these qualities is professional objectivity. One of the traditions that has been pounded into the head of everyone who emerges as a graduate scientist from a legitimate institution is that he must try to be objective in his approach. He must avoid depending upon authoritative opinions of others as his reasons for conclusions. He must scrutinize the raw data on which conclusions have been based. He must painstakingly check the logic at every stage of the game before he arrives at his own conclusion.

There are two admissions that must be made in connection with the objectivity of scientists. The first is that of course scientists have no monopoly on objectivity. Most good managers are pretty objective also. The second admission is that scientists don't always apply this professional objectivity to their politics or their private lives. However, this is by no means the only instance of a class of men who operate their business in accordance with very rigid professional rules and standards, but who do not always carry over such professional rules to their personal lives.

Scientists are no different than other people in these respects. The important thing is that in their professional lives they are heavily indoctrinated with the idea of objectivity. The standing that they have among their associates depends in a major way upon whether their work gives evidence of an unprejudiced, carefully substantiated objective approach. Most other classes of people are not subject to the same kinds of professional compulsions to be objective as are the scientists. As a consequence, on the average, the scientifically trained investigator of a business or industrial operation does a better job than others in separating fact from opinion and in arriving at conclusions that are properly related to the actual state of affairs under study.

QUANTITATIVENESS

A second attribute that is common to scientifically trained people is that of quantitativeness. The scientist is trained to work with magnitudes and not just with qualitative effects. The scientist typically is confronted with situations in which a variety of physical phenomena occur simultaneously. He learns at an early date that the proper approach to any new problem is first to determine which factors are the major ones and which are the minor ones, and

then to devote his principal attention to the important factors. A qualified scientist does not seize upon a narrow aspect of a problem that accounts for 10% of the phenomena he is investigating and concentrate his attention on that. He looks for the aspect of the situation that accounts for 90% of the phenomena and concentrates his attention upon that instead.

Such a quantitative approach is also not unique to scientists, but they have the advantage of having had this sort of thing drummed into them in their formal training and in their professional experiences. For years and years, they have spent most of their time in arriving at quantitative, numerical solutions to complex problems. Such a quantitative approach is of great importance in the analysis and resolution of operational and procedural problems that arise in a complex human society. The one thing that can usually be depended upon is that the problem that really is responsible for inefficient operations will be camouflaged by a large number of interrelated matters that can be differentiated from the major issue only by a quantitative evaluation of their relative effects on the attainment of the over-all objectives of the operation under study.

CAPACITY

Another pertinent quality of the scientist is his capacity for studying and learning new fields. In his day-to-day activities, the scientist has frequent need to acquire an understanding of the important elements of some new subject that he may never before have encountered. This capacity is the principal tool by means of which operations research scientists, whose past training may not have included business and industrial methods and procedures, have been able to demonstrate an ability to equip themselves with a good understanding of these nontechnical matters in what sometimes seems to non-scientists to be a remarkably short time.

There is one other advantage many scientists have that, when added to these other points, gives them an unusually favorable position for certain kinds of management consulting activities. As everyone today knows, we are in the early stages of what will within the next twenty years begin to look like a second industrial revolution, characterized by the wide-scale application to business and industry of automation techniques, both for the operation of factories and the handling of data processing assignments. The equipment that is being developed and applied to these tasks is, for the most part, of a highly complex character, and it is primarily electronic in nature. It just so happens that the broadening experience that so many scientists went through during the last war was accomplished in connection with the development of precisely these electronic techniques that are now beginning to be applied to business and industry. This was especially true of physical scientists and electronic research engineers who today comprise one of the most important new classes of professional entrants into the field of operations research. But it is in the application of some of these newer tools of automation and data-

processing that an important part of the growing need for operations research arises. While the manager of a company may or may not decide in favor of an operations research program in the normal course of his business, he has practically no alternative on the occasion of introducing the newer techniques of automation. The very nature of these new powerful tools frequently makes operational procedures simple that were completely impractical before, and at the same time practically requires the elimination of some procedures that in the past were reasonably efficient.

And so, not only does the objective quantitative approach that results from his professional training put the scientist in an unusually favorable position to be effective in operations research, but also the growing impact of automation techniques in business and industry places an unusual premium on the physical scientist who can deal effectively with the newer electronics techniques, as well as with generalized operation problems.

SUMMARY

By thinking of operations research as simply a quantitative objective approach to the solution of almost any complex management problem, and in minimizing the differences in meanings between the terms management consulting and operations research, the situation has been consciously and deliberately oversimplified. On the other hand, any generalizations have been much more nearly right than wrong, and by employing them it was hoped to succeed in clearing up some of the confusion that seems to surround this much discussed and little understood subject of operations research.

The principal points that have been made, in addition to the approximate synonymity of the term operations research with quantitative objective analysis of management problems, have been two in number. The first was that the thing that is new in this field of management consulting or operations research is that scientists have decided that such activities are respectable and, as a consequence, are getting into the field in large numbers.

The second point was that this is a good thing. As a class, scientists do a better job than most people of separating the major and minor factors in a complex operational situation and applying to them an objective quantitative analysis to arrive at recommendations for company procedures or decision-making methods that best serve the major over-all objectives of the organization. In the special class of situations which may be expected from now on to be the source of a steadily increasing fraction of operations research requirements, arising out of the application of automation techniques to business and industry, operations research groups containing physical scientists or electronic research engineers should be uniquely productive.

The scientist has now decided that the business man is a respectable partner. There are some indications that the business man is willing to be wooed and won. If a marriage does indeed result, there is every likelihood that the union will be a happy and fertile one.

✦✦✦✦✦✦✦✦✦✦✦✦✦ "OPERATIONS RESEARCH" FOR MANAGEMENT

CYRIL C. HERRMANN AND JOHN F. MAGEE

ESSENTIAL FEATURES

Operations research apparently means different things to different people. To some businessmen and scientists it means only the application of statistics and common sense to business problems. Indeed, one vice president of a leading company remarked that if his division heads did not practice it every day, they would not last long. To others it is just another and perhaps more comprehensive term for existing activities like market research, quality control, or industrial engineering. Some businessmen consider it a new sales or production gimmick; some, a product of academic people interfering in the practical world. In truth, operations research is none of these things, as we shall soon see. . . .

The first point to grasp is that operations research *is* what its name implies, research on operations. However, it involves a *particular* view of operations and, even more important, a *particular* kind of research.

Operations are considered as an entity. The subject matter studied is not the equipment used, nor the morale of the participants, nor the physical properties of the output; it is the combination of these in total, as an economic process. And operations so conceived are subject to analysis by the mental processes and the methodologies which we have come to associate with the research work of the physicist, the chemist, and the biologist—what has come to be called "the scientific method."

THE SCIENTIFIC METHOD

The basic premise underlying the scientific method is a simple and abiding faith in the rationality of nature, leading to the belief that phenomena have a cause. If phenomena do have a cause, it is the scientist's contention that by hard work the mechanism or system underlying the observed facts can be discovered. Once the mechanism is known, nature's secrets are known and can be used to the investigator's own best advantage.

The scientist knows that his analogue to nature will never be entirely perfect. But it must be *sufficiently* accurate to suit the particular purposes at hand; and, until it is, he must repeat the processes of observation, induction, and theory construction—again and again. Note that a satisfactory solution must be in quantitative terms in order that it can be predictive—the only accepted fundamental test of being physically meaningful.

The scientific method, in its ideal form, calls for a rather special mental attitude, foremost in which is a rever-

Harvard Business Review, *July–August 1953, 31:4, 100–112.*

ence for facts. Of course all modern executives are accustomed to using figures to control their operations. But they are primarily concerned with results and only secondarily with causes; they interpret their facts in the light of company objectives. This is a much different attitude from seeking out the relationships underlying the facts.

Thus, when an executive looks at sales figures, he looks at them primarily in terms of the success of his sales campaign and its effect on profits. By contrast, when the scientist looks at these same figures, he seeks in them a clue to the fundamental behavior pattern of the customers. By the process of induction he tentatively formulates a theoretical system or mechanism; then by the inverse process of deduction he determines what phenomena should take place and checks these against the observed facts. His test is simple: Does the assumed mechanism act enough like nature—or, more specifically in this case, does it produce quantitative data such as can be used for predicting how the customers will in fact behave? . . .

IMPLEMENTATION

Through the years mathematical and experimental techniques have been developed to implement this attitude. The application of the scientific attitude and the associated techniques to the study of operations, whether business, governmental, or military, is what is meant by operations research.

Newton was able to explain the apparently totally unrelated phenomena of planetary motion and objects falling on the earth by the simple unifying concept of gravity. This represented a tremendous step forward in helping men to understand and control the world about them. Again, more recently, the power of the scientific method was demonstrated by the ability of the nuclear physicists to predict the tremendous energy potential lying within the atom.

Here are a few summary examples of the way this same kind of approach has been applied to down-to-earth business problems:

A company with a number of products made at three different locations was concerned about the items to be produced at each location and the points at which the items would be warehoused. Freight costs constituted a substantial part of the delivered cost of the material. Operations research showed that what appeared to be a complex and involved problem could be broken into a series of rather simple components. Adaptations of linear programing methods were used to find the warehousing schedule which would minimize freight costs. The study is now being extended to determine the best distribution of products among manufacturing plants and warehouse locations in order to minimize net delivered cost in relation to return on investment.

A manufacturer of chemical products, with a wide and varied line, sought more rational or logical bases than the customary percentage of sales for distributing his limited advertising budget among products, some of which were growing, some stable, and others declining. An operations research study showed that advertising effectiveness was related to three simple characteristics, each of which could be estimated from existing sales data with satisfactory reliability: (a) the total market potential; (b) the rate of growth of sales; (c) the customer loss rate. A mathe-

matical formulation of these three characteristics provided a rational basis for distributing advertising and promotional effort.

In a company making a line of light machines, the executive board questioned the amount of money spent for missionary salesmen calling on customers. Studies yielded explicit mathematical statements of (a) the relation between the number of accounts called on and resulting sales volume and (b) the relation between sales costs and manufacturing and distribution costs. These were combined by the methods of differential calculus to set up simple tables for picking the level of promotion in each area which would maximize company net profits. The results showed that nearly a 50% increase in promotional activity was economically feasible and would yield substantial profits.

An industrial products manufacturer wanted to set time standards as a basis for costs and labor efficiency controls. The operations research group studied several complex operations; expressed the effect of the physical characteristics of products and equipment and the time required to produce a given amount of output in the form of mathematical equations; and then, without further extensive time study or special data collection, set up tables of production time standards according to product characteristics, equipment used, and worker efficiency, which could be applied to any or all of the production operations.

A company carrying an inventory of a large number of finished items had trouble maintaining sound and balanced stock levels. Despite careful attention and continued modification of reorder points in the light of experience, the stock of many individual items turned out to be either too high for sales or inadequate to meet demand. The problem was solved by a physical chemist who first collected data on the variables, such as size and frequency of order, length of production and delivery time, etc.; then set up an assumed system, which he tried out against extreme sales situations, continually changing its characteristics slightly until it met the necessary conditions—all on paper (a technique well known to physical scientists); and thus was able to determine a workable system without cost of installation and risk of possible failure.

These examples should serve to give some idea of how the scientific method can be applied. But they represent only a few of the many scientific techniques available (as we shall see when we examine further cases in more detail). Some practitioners even take the rather broad point of view that operations research should include the rather indefinite and qualitative methods of the social fields. Most professional opinion, however, favors the view that operations research is more restricted in meaning, limited to the quantitative methods and experimentally verifiable results of the physical sciences.

BASIC CONCEPTS

There are four concepts of fundamental importance to the practice of operations research: (a) the model, (b) the measure of effectiveness, (c) the necessity for decision, and (d) the role of experimentation.

THE MODEL

The most frequently encountered concept in operations research is that of the model—the simplified representation of an operation, containing only those aspects which are of primary importance to the problem under study.

It has been of great use in facilitating the investigation of operations. To illustrate with some familiar types of "models" from other fields:

(1) In aeronautical engineering the model of an aeroplane is used to investigate the aerodynamic properties in a wind tunnel. While perfectly adequate for this purpose, it would hardly do for practical use. It has no seats; it may not even be hollow. It is, however, a satisfactory physical model for studying the flight characteristics of the ship.

(2) Another, quite different kind of model, with which we are all familiar, is the accounting model. This is essentially a simplified representation on paper, in the form of accounts and ledgers, of the flow of goods and services through a business enterprise. It provides measures of the rate of flow, the values produced, and the performances achieved, and to that extent is useful (though it is hardly a realistic representation of *operations*).

(3) Many models are used in physics. Three-dimensional models of complex molecules are probably most familiar to laymen, but the most powerful models in this field are sets of mathematical equations.

There are several different types of operations research models. Most of them are mathematical in form, being a set of equations relating significant variables in the operation to the outcome. . . .

Another type of model frequently used is the punched-card model, where components of the operation are represented by individual punched cards; masses of these are manipulated on standard punched-card equipment. For example, in a study of a sales distribution problem, each customer, of thousands served by the company, was represented by a punched card containing significant information about his location, type of business, frequency of purchase, and average rate of business. The punched cards representing the customers could then be subjected to assumed promotional treatments, with the effects of the promotions punched into the cards. The resulting business could be calculated and an evaluation made of alternative sales-promotion campaigns.

Occasionally a model is physical like the ones often used by engineers. For example, the use of a hydrokinetic model has been proposed in the study of a mass advertising problem. The fluid flowing through the model would represent business of various types going to the company or to competitors as a result of various forms of the company's own and competitive promotional efforts (represented in the model by forces acting on the fluids).

Operations research models can also be distinguished as exact or probabilistic:

(1) An *exact* model is used in operations or processes where chance plays a small role, where the effect of a given action will be reasonably closely determined. Exact models can be used, for example, in long-range production scheduling problems in the face of known or committed demand. The exact model is sufficiently accurate since it can be assumed that, barring a major catastrophe, over the long run planned and actual production will be reasonably close.

(2) The *probabilistic* model, on the other hand, contains explicit recognition of uncertainty. Such models are of great use in the analysis of advertising prob-

lems, where the unpredictability of consumers plays a great role. . . . they make extensive use of the highly developed theory of probability, which has come to be of such great value in the physical sciences. One customarily thinks of a physicist as dealing with rather exact concepts and highly predictable experiments. Yet physicists faced a problem equivalent to the advertising problem in predicting atomic activity. Methods developed for physical problems involving mass behavior under random conditions can be applied with great facility and value to operations.

The model is a major goal of the operations research analyst. In one sense, the construction of the model, or a faithful representation of the operation, is the scientist's primary job. In doing it he develops a theory to explain the observed characteristics of the operation. . . . The remaining task is to interpret this theory through the manipulation of the model, whether mathematical or physical.

MEASURE OF EFFECTIVENESS

Related to the concept of a model or theory of operation is the measure of effectiveness, whereby the extent to which the operation is attaining its goal can be explicitly determined. One common over-all measure of effectiveness in industrial operations is return on investment; another is net dollar profit. Measures of effectiveness down the scale might be the number of customers serviced per hour, the ratio of productive to total hours of a machine operation, etc.

A *consistent* statement of the fundamental goals of the operation is essential to the mathematical logic of the model. (It does not matter if the goals are complex.) Just as the model cannot make 2 and 2 add up to 5, so it is impossible to relate fundamentally inconsistent objectives and produce consistent and meaningful results.

Operations research has frequently brought to light inconsistencies in company goals. Take production scheduling, for instance. Very often its object has been stated as scheduling production to meet sales forecasts with minimum production costs, with minimum inventory investment, and without customer-service failure. Yet minimizing inventory investment typically requires the use of start-and-stop or at best uneven production plans, resulting in excessive production costs; and eliminating the risk of not being able to ship every customer order immediately requires huge inventories, in the face of fluctuating and at least partially unpredictable demand.

The solution is to combine and sublimate such otherwise inconsistent goals to a higher unified and consistent goal. To illustrate:

> The diverse goals of customer service, production economy, and investment minimization can be expressed in terms of costs—the cost of inefficient production (hiring, training, overtime, etc.), the cost of investment in inventory (the rate of interest the treasurer wishes to charge to conserve his funds or perhaps the return on investment which can be earned through alternative uses of the available funds), and the cost of inability to meet a customer's demand (estimated loss of goodwill and future business). While the latter two costs are primarily policy costs, experience has shown that they are sufficiently determinable and realistic to afford a basis for management decision.

The three component costs can then be cast in an algebraic equation expressing their interrelationships in terms of total scheduling cost; and the minimum total scheduling cost becomes the one, consistent goal.

Note that, once set up, the algebraic equation can be worked in reverse. Thus, the sales manager might be told how much the company can *afford* to pay for an inventory large enough to avoid varying risks of failure to meet consumer demand.

This kind of clarification of goals is particularly important in relating subordinate and over-all company goals—as in the case of a department run efficiently at the expense of other departments or of a promotion budget based on a fixed percentage of sales without regard to the adverse effects on manufacturing budgets.

The statement of a complete and wholly consistent goal of company operations must be recognized as an ideal. Business goals are very complex, and to catch the full flavor of the objectives of an intricate business operation in any simple, explicit statement is difficult. Many business goals remain, and probably ever will remain, at least in part intangible—e.g., efforts to improve employee morale or contribute to the public welfare. To that extent, the objective of operations research must be more modest than the construction of a complete model and the measurement of the extent to which the operation is attaining the complete set of goals established for it. But it still can serve to clarify the interdependency of those intangibles with the company goals which in fact are measurable, thus providing a guide to executive decision.

NECESSITY FOR DECISION

The third concept inherent in operations research is that of decision and decision making. An essential element in all true operations research problems is the existence of alternative courses of action, with a choice to be made among them; otherwise the study of an operation becomes academic or theoretical. This should be clear from the cases already cited.

In sum, the objective of operations research is to clarify the relation between the several courses of action, determine their outcomes, and indicate which measures up best in terms of the company goal. But note that, while this should be of assistance to the executive in making his decision intelligently, in every case the ultimate responsibility still lies with him.

ROLE OF EXPERIMENTATION

The fourth significant concept concerns the role of experimentation. Operations research is the application of experimental science to the study of operations. The theory, or model, is generally built up from observed data or experience, although in some cases the model development may depend heavily on external or a priori information. In any event, the theory describing the operation must always be verifiable experimentally.

Two kinds of experiments are important in this connection:

(1) The first kind is designed simply to get information. Thus, it often takes the form of an apparently rather impractical

test. In one case the operations analysts directed advertising toward potential customers the company knew were not worth addressing, and refrained from addressing customers the company typically sought—and for a very simple reason. There was plenty of evidence indicating what happened when advertising was directed toward those normally addressed but not enough about its effects upon those *not* normally addressed. To evaluate the effectiveness of the advertising, therefore, it was necessary to find out what happened to those normally promoted when they were not promoted, and what happened to those normally not promoted when they were.

(2) The other type of experiment is the critical type; it is designed to test the validity of conclusions. Again, what appear to be rather impractical forms of experimentation are sometimes used. Thus, in the most sensitive experiments of this type, the validity of the theory or model can often be tested most revealingly in terms of the results of extreme policies rather than in terms of the more normal policy likely to be put into practice.

OTHER SERVICES

Now, before going on to discuss in more detail the administrative problems and uses of operations research, it may be well to make clear how it differs from other services to management. Many of these services have been proved of great value to the business community as a result of years of successful application to difficult problems. Are there significant differences that make it possible for operations research to extend the usefulness of these services? Let us examine some of the leading services briefly for comparison:

Statistics. Operations research is frequently confused with statistics, especially as applied to the body of specific techniques based upon probability theory which has grown up in recent years. This statistical approach originally developed in the fields of agriculture and biology but has now been extended into such areas as quality control, accounting, consumer sampling, and opinion polls.

The operations research analyst does use such statistical methods when applicable, but he is not restricted to them. Moreover, there is a difference in basic point of view. Statistics is concerned primarily with the relations between numbers, while operations research is concerned with reaching an understanding of the operation—of the underlying physical system which the numbers represent. And this may make a significant difference in results as well as approach. In a recent advertising study, the operations research team found the key to characterizing the way in which the advertising affected consumers in the results of a series of "split-run" tests. Earlier, these results had been presumed useless after statistical methods such as analysis of variance and multiple regression had failed to show meaningful conclusions.

Accounting. Operations research is also confused sometimes with accounting, particularly with the control aspects of accounting which have developed in recent years. In reality there are several differences. One springs from the fact that the fundamental and historical purpose of accounting methods has been to maintain a record of the financial operations of the company; and this is reflected in the training and attitude of many accountants. The growth of the accounting function as the interpreter of information for control purposes has been a fairly recent development, and the basic methods used and information provided are strongly influenced

by the historical accounting purpose.

Accounting information is one of the principal sources of data to support an operations research study. Accounting data, however, require careful interpretation and organization before they can be used safely and efficiently. Businessmen tend to forget that accounting costs are definitions derived in the light of the fundamental accounting purpose, and sometimes they tend to confuse accounting figures with "truth." Operations research, using the same raw data, may make other definitions which serve the special needs of the particular study. One of the great stumbling blocks in the organization and implementation of an operations research study is the disentangling from accounting records of the costs appropriately defined and truly significant to the problem at hand.

It is true, however, that in the analysis and construction of measures of control, the functions of operations research and accounting do tend to overlap. Also, the men working in these functions have strong mutual interests. Accountants have served a useful purpose in bringing the importance of control measures to the attention of business management, while operations research has shown ability in building new methods for developing and implementing these concepts of control.

Marketing Research. This management service is concerned with gathering and analyzing information bearing on marketing problems. Certain marketing researchers do go so far that in some instances they are performing services akin to operations research, but for the most part they are content to measure the market, by the use of questionnaires, interviews, or otherwise, and to gather factual data which management can use as it sees fit.

By contrast, operations research, when applied to marketing problems, seeks to gain a greater understanding of the marketing operation rather than of the market itself. Thus, it may rely heavily on marketing research sources for data; in one retail advertising study, for example, a consumer-interview program was used to obtain information on the frequency with which potential customers purchased outside their own towns. But the objective, even in quantitative studies, is usually to obtain a fundamental characterization of consumers for use in the model. Furthermore, much of operations research in marketing problems is directed toward clarifying the interdependencies between marketing and other company operations. Finally, it draws on a range of techniques and analytical methods that are well beyond the scope of the usual marketing research.

Engineering. Again, the boundary between operations research and engineering is frequently unclear. Some examples may serve to draw it more definitively:

(1) During the last war a great deal of effort went into the improvement of the effectiveness and efficiency of depth charges. The objective of engineering and physics research was the construction of a depth charge having the strongest explosive power. Operations research, however, was concerned with the effective use of the depth charges then available for the purpose of sinking submarines.

(2) In a recent industrial situation, the engineering problem was to construct a new railway control system which would get control information quickly and clearly to the railway engineer. By contrast, the associated operations research problem was to determine whether increased speed and clarity of control information would help the train engineer in his task of getting the train to its destination safely and quickly.

(3) More subtle distinctions can be found in the study of equipment that

tends to break down in operation, such as aircraft or chemical-process equipment. The engineering problem may be to find out why the equipment breaks down and how the breakdowns can be prevented. The operations research assignment is likely to be finding the best way to run the operation in view of available information on the relation between breakdown and use.

Industrial Engineering. Perhaps the most difficult distinction to make is that between operations research and modern industrial engineering. The pioneers in the field of industrial engineering did work of a character which operations research analysts would be proud to claim for their field.

In modern practice, however, industrial engineers usually apply established methodologies to their problems. Moreover, their work is generally restricted in scope to manufacturing activities and, in some cases, to distribution operations. Equally important, industrial engineering is not commonly characterized by the mental discipline and techniques of analysis that are commonly associated with the physical scientist; operations research is.

Perhaps the most significant difference marking off operations research from other management services lies in the type of people employed. Operations research people are scientists, not experts. Their value is not in their knowledge or business experience but rather in their attitude and methodology. It is indicative of the influence which the physical sciences have exerted on the people in operations research that they have a self-conscious concern with concepts and first principles and show a desire to generalize from specific examples to all-encompassing theories.

In any event, the important point is that, far from supplanting or competing with other management services, operations research has been shown by experience to be particularly successful in those areas where other services are active and well developed. Indeed, one useful contribution of operations research is frequently that of integrating other information, of using the expert opinion and factual data provided by other services in an organized, comprehensive, and systematic analysis. A soundly organized operations research group should have available the services and counsel of experts in these fields for most effective joint attack on management problems. For example:

> In the continuing research program of one retail store chain operation, marketing research methods are used to provide field observations, opinions, and data on the behavior of consumers.
>
> The accounting organization provides information on costs and capital requirements.
>
> In the operations research models these data are combined and interpreted to yield information on cost control, staff incentives, merchandise policies, and credit management. . . .

EVALUATION

In perspective, what is the current status of operations research? What are its contributions, its limitations, its future?

CONTRIBUTIONS

Case histories show that operations research provides a basis for arriving at an integrated and objective analysis of operating problems. Characteristi-

cally, operations research tends to force an expansion in viewpoint and a more critical, questioning attitude. It also stimulates objective thinking, partly because it emphasizes broad purposes and partly because the mathematical nature of the model and techniques limits the influence of personal bias.

The results of operations research studies are quantitative. They provide an opportunity for sound estimates in terms of requirements, objectives, and goals, and a basis for more precise planning and decision making.

The contributions of operations research to business analysis and planning have been important and substantial. Here are two worth singling out:

1. *The application of organized thinking to data already existing within the company*—Frequently a major contribution has been the location, collection, and classification of existing data scattered through widely separated branches of the company. In one recent study, an operations research team found the same fundamental problem cropping up under various guises in a number of different parts of the company. Each division or section had its own point of view toward the problem, and each had significant information bearing on it that was unavailable to the others. This sort of thing happens despite the most sound and progressive management; operations research tends to rectify it.

2. *The introduction of new concepts and new methods of analysis*—Some of these concepts, such as information theory, control theory, and certain aspects of statistical mechanics have been carried over from other fields; the physical sciences, and in particular modern physics, have been a very fruitful source of transplanted analytical techniques. But there are also certain original contributions, such as the newborn theories of clerical organization and consumer behavior, which suggest the possibility of developing further tools for attacking important business problems. All these techniques make it possible to explore the effects of alternate courses of action before management becomes committed to one of them.

LIMITATIONS

Operations research is hardly a cure-all for every business ill; neither is it a source of automatic decisions. It is limited to the study of tangible, measurable factors. The many important factors affecting business decisions that remain intangible or qualitative must continue to be evaluated on the basis of executive judgment and intuition. Often they make it necessary to adjust or modify the conclusions drawn from the quantitative analysis of the researchers. Professional personnel in operations research strongly emphasize this distinction between the operations research responsibility for analysis and the executive responsibility for decision. They point with approval to cases like this one:

> In a recent series of conferences called to implement the results of a long and major operations research investigation, the analysts emphasized that their conclusions were based in part on the assumption that the output of a plant in question could be increased substantially at the existing level of efficiency. The executive responsible for the operation of the plant felt that this assumption was a sound one. The official responsible for the ultimate

decision, however, decided to follow a more conservative course of action than the one indicated by the study, primarily because of his estimate of the psychological effect that increases in volume would have on the plant personnel.

The fact that operations research is scientific in character rather than expert means that more time is required to achieve useful conclusions than in the case of normal engineering analyses. As an applied science, the work is torn between two objectives: as "applied" it strives for practical and useful work; as "science" it seeks increasing understanding of the basic operation, even when the usefulness of this information is not immediately clear. The executive who plans to support research work of this character must be fairly warned of the need for restraint. The natural tendency to require that the studies or analyses be "practical" can, if enforced too rigidly, result in the loss of substantial benefits. Also, the results of studies of this type are necessarily somewhat speculative. When operations research is purchased, neither the specific program to be followed, the precise questions to be answered, nor the successful achievement of results can be guaranteed.

Recognition of this difference between operations research and more conventional engineering methods is essential to the satisfaction of both the controlling executive and the analyst. . . .

NEW HORIZONS

In conclusion, the future of operations research appears reasonably bright at the present time. Successful applications in industry are fulfilling the hopes of its early supporters, and the skepticism of businessmen is tending to break down as successful case histories pile up and become available for publication.

The areas of potential application of operations research appear broad. The future holds possible extensions such as the development of strategic concepts through the applications of the much heralded (but as yet largely untested) theory of games and by the development of a fundamental understanding of the impact of advertising and merchandising methods.

How will operations research help in the future to clarify the role of the executive? Present indications are that it will live up to its expectations of helping executives to make decisions more intelligently, but the decisions will always remain to be made. The possibility of removing all subjective and qualitative factors must be deemed at the present time to be more a hope than a real possibility, and the construction of completely consistent and logical goals, while a reasonable objective in decision making, is probably unattainable. The balancing of the responsibilities to society, consumers, owners, and employees will therefore still be the fundamental task of executives.

✦✦✦✦✦✦✦✦✦✦✦✦✦ BASIC DEVELOPMENTS THAT MAKE OPERATIONS RESEARCH AND SYNTHESIS POSSIBLE

MELVIN L. HURNI

. . . three important developments . . . have made possible the emergence of Operations Research & Synthesis. These are developments that have come out of the study of business as a phenomenon rather than out of the practice of business. The origins of these developments are not recent. They informed Frederick W. Taylor in his work, even though he may not have specifically stated them. They are inferred in the monumental writings of Harry Arthur Hopf. They received expanding if perhaps accidental recognition and use during the more recent war years. The evidence at hand in respect to their reality is sufficient to make the business executive who will take the time to inspect it, ponder if there is not available a new opportunity to find out more about what goes on and why in his business—to the end of strengthening and bringing greater assurance to his decision making.

These developments are as follows:

The *first* is a seemingly simple insight that business management is not just the result of feel, intuition or experience that results in inspired decisions—that the executive himself cannot explain, but that it is to a very significant extent the result of rational action. In other words, even though the business executive may view a market or any other situation as uncertain or fleeting, or not completely known, when he decides to take action for whatever cause, he proceeds on a rational basis.

By this is meant that the manager focuses on specific objectives, be they the reducing of price levels, the changing of production schedules, or expanding facilities. He bases his thinking on assumptions regarding the environment and the resources available. He appraises risks and weighs them against attainable benefits. He attempts to identify alternative courses of action and selects one of them that to him seems to offer the most favorable balance among effort, risk and likely result. He has expectations regarding the outcome of the course of action chosen, and these expectations establish a basis for measuring results and for revising and changing the decision made if circumstances change, or if the results prove the decision inappropriate.

A manager may now have to "guess" or "play hunches" with respect to every one of these elements. In fact, there may always be a degree of uncertainty or

"The Purpose of Operations Research and Synthesis in Modern Business," *Management Consultation Services of the General Electric Company, 1955, 2–7.*

irreducible ignorance in making decisions. However, if he reasons in this manner and is not acting on inspiration, it should be feasible to bring to these notions of the nature of the business, the market, the resources and the effect of action, more precision and significant detail through systematic study of each, not merely when decision is to be reached, but on a continuing basis.

The *second* development is that systematic study of these elements in a number of business situations has already brought to light a basic and significant orderliness in an increasing and expanding range of business phenomena. In other words, business and economic life are not entirely haphazard as many of us would like to believe. On the contrary, basic patterns of order underlie many business phenomena.

It must be admitted that we are not dealing with immutable patterns of order such as are disclosed by scientific research in the physical world, but with dynamic and shifting patterns. It is significant to note, however, that these patterns do have reasonable life span, and hence may be utilized for the attainment of economic or social purposes.

Although the idea of rational action may come easily to the experienced business executive, the notion of a basic orderliness of some life span for business phenomena may be more difficult to accept. Yet we admit of this orderliness, even though we do not describe it when we build a new plant and expect the basic requirements will remain sufficiently unchanged to permit pay-off in from 5 to 20 years. We redesign our products and retool with the same expectancy of a basic orderliness of sufficient duration to assume pay-off. We expect the ultimate in this respect when we go to automation.

Similarly, we promote people on the basis of their past experience and performance. This, too, presumes that what has happened in the past has significant validity for the future.

Perhaps we do instinctively recognize these patterns of order without being too aware of them. What is important is that we not only develop awareness, but also give them definition not only in respect to their content, but in respect to how dynamic or how shifting they are in terms of the basic processes of our particular business. With such knowledge, we can plan and decide with greater assurance for both present and future.

The discovery of patterns and relationships already covers a great range of situations for a particular business. An examination of the literature will disclose investigation of such things as the relation of cost and volume, volume and price, machine failure and volume, profitability and product mix, to mention a few. Here again the objective is a crisp and understandable description that can be communicated, not just a feel or sense for the situation resting on the judgment of a single individual.

One senses a close kinship between these kinds of relationships and the characteristics and properties of materials and mechanisms familiar to the physical scientist and engineer. They should permit the executive and his organization to understand the situa-

tion more clearly so that they may maximize or balance among elements for optimal performance as situations arise.

The *third* development is the growing recognition of the applicability of methods of investigation out of the physical sciences and out of mathematics and logic. This makes it possible to define and describe business phenomena and managerial problems in simple and often in quantitative form. This in turn makes possible systematic analyses of situations, intelligent anticipation of consequences through the synthesis of hypotheses or models of situations, a high degree of measurability and also of communicability.

It is these three developments that are of significance and must be grasped by business executives who wish to understand or apply Operations Research & Synthesis. For Operations Research & Synthesis is founded on them and is informed by them. It is not a fire fighting procedure, a means for solving spot problems, or a collection of canned methods. It is analogous in an important sense to physical research as conducted in industrial laboratories in its processes, methods and outlook. Without the increasing evidence and recognition of these three developments it would not be possible at all.

WHAT THE BUSINESS EXECUTIVE MAY EXPECT FROM OPERATIONS RESEARCH & SYNTHESIS

The major aim of Operations Research & Synthesis is to disclose and strengthen the rational system underlying the business enterprise. It does this principally through disclosing facts, relationships and characteristics, wherever possible, to replace opinion, vague generalities, or lore. It does this further, as the lore of the business is replaced with fact, through developing hypotheses, models or analogies that describe how parts of the business and the whole business fit together and work. It thus provides a basis for the prediction of performance or consequences, at least within the range of expected conditions. It thus provides means for not only developing better understanding of the system in operation, but also a means of foreseeing and pre-testing possible improvements.

To many experienced business executives, this latter aim is by no means new. The simple fact that in most businesses the facilities are operated by organizations of people requires that there be some measure of clarity in respect to these elements in order that the enterprise may have some semblance of unity and coherence of purpose. The familiar and currently popular work in respect to organization structuring, position design and description, reservation and also delegation of authority and responsibility, and policy making are a part of this aim.

What is unique about Operations Research & Synthesis is the organization of this work around the purposeful and directed application of the scientific approach which brings with it, from its very nature, the expectation of greater precision and detail in description, the possibility of testing over ranges of circumstances, the basis for predictability of performance, measurement and communication.

We might then expect from Operations Research & Synthesis work four specific things:

The *first* is knowledge of how things behave under a range of stated conditions and in increasingly precise form.

The *second* is tested ideas why they behave in this manner, or tested ideas how this behavior may be utilized to meet specific objectives, and how well.

The *third* is tested ideas how this behavior might be changed, consistent with specific business objectives.

The *fourth* is increasing insight in respect to the nature of the business and its characteristics.

This last expectation is worthy of further explanation by means of an example. One such example is from a major railroad which began its work in Operations Research & Synthesis with the study of individual problems. It studied paper work procedures, such as freight billing, in view of a managerial objective to cut costs. It studied operating problems, such as freight car allocation, for the purpose of relieving bottlenecks. It studied individual capital investment decisions, such as the capacity of a new switching yard. In all these problems, Operations Research & Synthesis provided information that enabled the managers to make significant improvements in the mode of operations or to make better decisions.

But at the same time, these particular studies and others like them led to a general insight regarding the nature of the business and its major characteristics and to the kinds of major changes required to relieve such symptomatic and continuing problems as the aforementioned high costs and bottlenecks. It led to a *hypothesis of the business;* namely, to the hypothesis that a railroad in its basic economics is a process rather than, as had always been assumed, a series of individual job-shop operations. This hypothesis implied that the basic economic problems of a railroad are the rate at which it utilizes its capital equipment and the "yield-mix" between different kinds of "products" put through the total system.

This insight in turn led to significant improvements and to significantly better decisions. But at the same time, it also indicated the need for further research in specific functional areas. One of these areas was that of organization, particularly the question whether the traditional "building blocks" of railroad organization structure were really adequate. Another study area that came out was that of the proper classification of the business of the railroad in respect to economic characteristics and profitability, aimed at the development of a purposeful marketing plan based on an optimal "product mix."

Another area of study found was that of equipment design, equipment capacity and equipment location in light of this new-found knowledge. And while each of these studies is in itself a long-term project which will not yield results for several years, they have already yielded further insight into the basic hypothesis of the business and have made thereby possible further sharpening of managerial decisions in all areas and on all levels.

We thus find Operations Research & Synthesis work performing three major functions:

The *first,* transforming lore about the

business into knowledge, as for example, describing incoming orders for certain types of businesses not as discrete events, but as a continuing activity in time having a typical rate and an expected variation about this rate.

The *second,* replacing feel or intuition about what goes on within the business, its markets and its environment by tested or testable knowledge, as for example, the relationship between all parts and subassemblies to all finished models of product.

The *third,* providing insights that help the manager to develop a rational and systematic hypothesis regarding his entire business, which enables him to integrate individual functions and specialized operations with the whole business and the whole business with the economic process of which it is a part, as seen in the earlier railroad example.

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE NEEDS AND OPPORTUNITIES FOR OPERATIONS RESEARCH AND SYNTHESIS

MELVIN L. HURNI

The need for Operations Research & Synthesis has been with us for a long time. Judging by the *needs* of . . . a large company and of its managers, Operations Research & Synthesis can be said to have been slow, if not tardy in its arrival. For managers have long been faced with the problems of technological and economic complexity; with the need to provide unity of direction, vision and effort while yet obtaining the benefits of functional specialization; and with the diversity and complexity of timing and time-dimensions in the modern business.

These needs have of course not gone unsatisfied. Managers have long and consistently developed improved and effective methods and tools to enable them to manage rationally and professionally. The organization structure as well as its management philosophy have been evolved specifically to strengthen both the ability to make decisions, and the authority to make and to implement decisions, despite the emergent size and complexity of our business.

Operations Research & Synthesis is therefore but a logical evolution rather than a radical innovation. What is new is perhaps only the ability to do in organized and systematic fashion what hitherto had to be done piecemeal or sporadically. This new ability, however, did not exist until recent years. Operations Research & Synthesis therefore not only answers a need of managers, it provides a significant *opportunity* as well.

"Operations Research and Synthesis in General Electric," *Management Consultation Services of the General Electric Company, 1954, 28–33.*

A. THE NEED

The need which Operations Research & Synthesis aims to satisfy arises out of such conditions as:

1. *The Increasing Complexity of Business Operations*—This results from growth, product diversity, technological developments, the lengthening futurity of business decisions, the impact of new and complex laws and governmental regulations, and from competitive as well as social pressures. These conditions point to a need for an even higher degree of precision in decision-making and implementation.

There thus is increasingly needed a growing body of analytical, logical and conceptual skills to establish realistic alternative solutions, to express key factors in such solutions, and to develop measures of their effectiveness, so that:

a. The risks inherent in a course of action can be judged rationally, prudently and, to the greatest feasible degree, in advance of the time when fundamental decisions must be taken.
b. The information needed by others for implementation may be precisely stated.

2. *Automation*—The tendency toward increasing the automatic machine content of factory and distributive facilities and of office equipment calls for an increasingly higher order of quantitative integration of both primary functional as well as subfunctional classifications of work. The assumption that the manager, from his knowledge, experience, perception alone and the understanding of the need to work with others, can integrate his activities and those of all applicable functions to the common good may progressively become invalid under such conditions. If nothing else, closer tolerances may be required than can be obtained by experience or perception alone; or, in other words, relationship responsibilities become as vital as purely functional responsibilities, and "teamwork" becomes as completely essential as "work" by the individual alone. And the more the production and distribution processes are made "automatic" the more such teamwork among the men performing specific functional work needs itself to take on the characteristics or at least the more precise relationships of automaticity.

There is required, increasingly therefore, a properly organized body of facts out of which quantitative relations have been distilled and evaluated to the point where a manageable body of common knowledge is available so that common purposes and common interests may be discerned and heeded. Only then can experience, managerial judgment, and good will be successfully applied to such matters as:

a. Appraising the need for and the relative economic effectiveness of various degrees and types of automaticity,
b. Surveying the expected conditions which must be met by automatic equipment; evaluating the degree of probability that the expected conditions will actually occur; and assuming the risks which the installation of facilities of a particular type may impose on other areas of the business,

c. Determining the specific kinds of information to be exchanged among functions and subfunctions for effective managing of automatic facilities,
d. Delineating the framework to provide over-all objectives and also to define the nature and extent of freedom for action in functional and subfunctional areas under Automation.

3. *Business as a Flow Process*—Slowly evolving understanding is growing that any business, viewed in proper perspective, is a flow process, if not in physical things, then in terms of information.

This poses a problem of integration similar to that posed by Automation. It points up that the need for integration comes not only from within but from without the business entity. This flow is *really* not broken by institutional, organization or ownership subdivisions but in fact extends from end purchaser back through the business chain to raw materials. Where this interdependence is not recognized, institutional, organizational, and ownership barriers tend too easily to establish points of distortion in such fundamental flow processes.

Here as in the previous case, the need for more precise knowledge in order to integrate within the whole chain is clear. The need is even more basic, since common interest among components within the chain, may not be apparent as they are within a component itself; and as free exchange of relevant information may be even more difficult. Common purpose and direction beyond a component or even a departmental business arising out of having common owners, servicing common customers and buying from common vendors may be as crucial in a given situation as common purpose and direction within a component. But, it may not be visualized easily, or even be readily capable of accomplishment at all by traditional methods. Where diverse ownerships of components in the chain are present, the requirements for an importance of sensing the need for over-all flow and integration is still more complicated and difficult to meet.

4. *The Manager's Needs*—It may not even be enough for a manager to know a product or group of products. Increasingly he may have to know, understand, and consider events outside of his field or market, on a national, if not on an international scale; for instance, in planning his supply of raw materials or in considering future population trends or long-range social philosophy in planning, for example, a plant lay-out.

For these reasons there is required by the Manager *a study of the environment* in which a particular business entity exists, the identification of uniquely relevant phenomena and relationships; the statement of relative stability of such phenomena; the evaluation of the relative expectancy of each; and the establishment of the methods by which the Manager may subsequently measure the tendencies of such phenomena and thereby more adequately inform his judgment and supplement his intuition and his past personal experience.

B. THE OPPORTUNITY

The opportunity to meet these needs is founded in a growing awareness that many business situations have a marked similarity to those in the natural sci-

ences that have yielded to definition within the last half century. In particular, the following major problems facing the manager of a business, parallel closely basic problems of the scientist. . . .

1. *Time as a Part of the Process*—The business manager faces a peculiar problem with respect to time. Time is not just a dimension in which phenomena happen. It is in itself a part of the process.

Changes in the business situation are apt to be imperceptible but cumulative; that is, they are likely to build up by a process of relatively slow "creep," or drift, from the familiar and established patterns on which past experience and current policies and procedures have been based. They are apt, therefore, to result, apparently suddenly, at some moment in time in a complete change in the basic situation, like the famous collapse of the durable One-Hoss Shay. Conventional methods, for instance, the normal methods of analyzing and reporting current operating information, cannot predict when this moment will occur, just as measuring the run-off on the surface of an iceberg cannot predict when it will turn turtle. Only determination of the rate of change of the center of gravity can do that.

On the basis of the conventional approach, these changes appear, therefore, to be unpredictable while at the same time the period between such changes appears to be unchanging, if not timeless. Operations Research & Synthesis should make possible the determination of the nature of such changes and the anticipation or prediction of the time at which the cumulative effect will bring about a basic regrouping, and with it a new situation.

2. *Measurements and Their Inter-relations*—The business manager faces a problem of Measurement in that measurements are relative to each other rather than absolute. The manager therefore has to decide what measurements are applicable and relevant and how to relate the readings of different measurements to each other. This closely parallels the measuring tasks for which the scientists developed some of the specific approaches and methods used in Operations Research & Synthesis.

3. *The Multi-dimensional Nature of Business Problems*—Perhaps a majority of problems are commonly multi-dimensional rather than single dimensioned. Therefore, they require for their analysis some kind of mathematical model which relates the data to other known information. Only by using such a model can the manager predict outcomes, starting with known or assumed conditions. The methods of Operations Research & Synthesis are basically methods of building logical models presenting multi-dimensional situations, preferably in quantifiable form.

4. *The Probabilistic Nature of Business Events*—Typically in a business situation the manager cannot say that a given event will always follow another. The best he can do is to say that there is a greater expectancy that this event, rather than another, will follow. He therefore needs a quantitative measure of the *likelihood* of various events occurring. Until this is done, data by themselves in many situations tend to

be more of a curiosity than a workable body of fact.

5. *Risk-taking*—Similarly, in business the manager is faced with alternative courses of action. Every business decision, therefore, is a risk-taking decision, which chooses between alternatives rather than finds one complete and exclusive answer. The professional manager, therefore, requires an approach that will identify the existing alternatives, will enable him to assess the risk ratio, and will help him to select, and at the time his decision is required, that alternative which minimizes the expected loss while maximizing the expected gain.

6. *Continuity and discontinuity*—Finally, business problems rarely are of the nature of a continuous "function" in the sense, for example, of the engineering or the mathematical function. The data with which the business manager deals are more likely to be aggregates of functions, quite frequently containing discontinuous elements as well as continuous ones. In many instances, it is not possible to discern by inspection any degree of continuity whatsoever.

In addition, this situation is likely to contain not just a few but many variables that affect the end results. In such a situation neither intuitive methods nor conventional business records of past events will necessarily lead to effective or valid results. However, the problem is extremely similar to problems in biological research from which some of the methods of Operations Research & Synthesis were specifically developed.

The *need* for the systematic and organized method that is Operations Research & Synthesis is particularly great in view of the nature of a company's business, its size, its technological complexity, the time span for which decisions have to be made and the goals business has set itself. The *opportunity* is however equally great. . . .

IN SUMMARY

Judging by the *needs* of the large, modern company and of its managers, Operations Research & Synthesis can be said to have been slow, if not tardy in its arrival.

Operations Research & Synthesis also presents significant *opportunities* to a company and its managers.

Some of the *needs* which Operations Research & Synthesis aims to satisfy are those arising from:

the increasing complexity of business operations
automation
the character of business as a flow process
the manager's needs for knowledge and information.

Basic characteristics of business operations offer specific *opportunities* for Operations Research & Synthesis—because

time is a part of the business process.
measurements and their inter-relations have to be defined.
many business problems are multi-dimensional.
business events are typically probabilistic.
risks have to be taken.
business problems are rarely a "continuous" function.

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE BASIC PROCESSES OF OPERATIONS RESEARCH AND SYNTHESIS

MELVIN L. HURNI

[Operations Research] consists of three related types of activity:

A. JUDGMENT PHASE—Defining the situation so that an attack of proper breadth and scope may be initiated;

B. RESEARCH AND SYNTHESIS PHASE—Describing and explaining the features and relations existing among the pertinent facts so that they may be purposefully employed;

C. ACTION PHASE—Communicating the results in such a manner that they may form a basis for initial and continuing managerial judgment and action.

A more detailed statement of these phases follows.

A. THE JUDGMENT PHASE

There are three major steps in the Judgment Phase.

STEP ONE—DEFINING THE FRAME OF REFERENCE

It is essential first to discover the nature of the situation. To draw an analogy from the natural sciences, it is desirable to know if a situation is in the field of chemistry, physics, or perhaps biology before any further activity is started. It is also essential to discover whether there will be, or will need to be, a fine laboratory in which to conduct the work or merely the drain board of a kitchen sink.

Among matters of relevance in this area are, for example:

1. *Basic assumptions* with respect to the situation:
   a. Determination of the kind of situation—whether manufacturing, engineering, marketing, general managerial or other areas;
   b. Determination of the degree of boldness the responsible managers will accept in its solution—do they desire to adapt more closely to existing circumstances or will they modify by reaching beyond these circumstances to new possibilities;
   c. Determination of the range of calculated risks which such managers are willing to take;
   d. Classification of the situation as an operational problem within an existing process or as one requiring modification of the process;
   e. Expected impact of the resulting decisions upon other areas or the whole business.

"Operations Research and Synthesis in General Electric," *Management Consultation Services of the General Electric Company, 1954, 11–26.*

2. The *characteristics* of the situation under study:
   a. Its recurrence—Is it:
      1) An isolated event?
      2) A recurring situation calling for the development of a methodology of some life span?
      3) A situation which, whether isolated or recurrent, is interlocked with others?
   b. Current knowledge with respect to the situation—its scope, information content, and consistency;
   c. Limits of time within which a decision must be made.
3. The *necessity for reversibility* in the resultant process:

For example, a conclusion to use specialized automatic machinery in the manufacturing activity, justified at certain levels of volume, might be costly to reverse. Hence, expectations of lower volume at some time in the not too distant future might make this decision untenable because of the cost and difficulty of going back to a more flexible system.

Other considerations might well occur in respect to the definition of the frame of reference of the situation. Those cited have been intended merely to give an indication of what is meant by "frame of reference."

STEP TWO—FINDING THE OPERATIONAL CHARACTERISTICS APPROPRIATE TO THE SITUATION

In addition to a knowledge of the framework in which the problem is set, it is essential to have an understanding of the type of operational mechanics required to provide the expected results. An appraisal of these requires consideration of such factors as:

1. The purpose of the activity—Is the purpose, for example, to set up a structure to give information only or a structure to make possible decisions that are consistent, balanced, intelligible, and effective throughout the organization;
2. The degree of accuracy required in the results;
3. The degree of measurability in the situation;
   a. Adaptable to exact measure;
   b. Adaptable to measurement with a margin of error;
   c. Adaptable to measurement with a probability distribution;
   d. Discontinuous but recurrent phenomena.

Indications like these provide a knowledge of the kind of data required, of the extent to which mathematical tools may have to be employed, and of the kind of mathematical skills that may be needed.

STEP THREE—FEEDING BACK ACQUIRED INFORMATION

The work of defining the situation may result in modification of what was originally presumed to be within the frame of reference or in the operational characteristics. This process of "feeding back" newly acquired knowledge and of modifying the process accordingly continues during the entire study of the situation. This feeding back and modifying is applicable to other phases of the activity to be described later as well as those already described. Among other things, this is a method by which

the activity moves from the study of symptoms to the study of causes.

B. THE RESEARCH AND SYNTHESIS PHASE

There are six steps in the Research and Synthesis Phase.

STEP ONE—DETERMINING THE METHODS AND UNITS OF MEASUREMENT

Determination of the methods of measurement of data will normally be a necessary first step in this phase. *Data commonly available may not be adequate either in form or nature for the purpose; or the data may be expressed through and in units of measurement that are not appropriate.* For instance, realization percentage of a budget of dollars of manufacturing cost may express the over-all performance of a shop, but it does not give a detailed indication of shop performance by classes of product or details of schedules, nor does it indicate the probable causes of under or over performance. . . . Determining the methods of measurement accordingly includes:

1. *Determining What is Pertinent,* as for example
   a. Isolation of the processes or activities involved in the situation;
   b. Statement of the desired or intended purpose of the activities or processes;
   c. Description of the factors that depict the degree of accomplishment or movement toward the accomplishment or intended purposes.
2. *Establishment of Units of Measurement*
   Establishment of the relationships to facts or between facts that cause to be displayed changes of magnitude, frequency, constancy, variability, probability, etc. of the pertinent facts, and the decision therefrom what units of measure are appropriate.

STEP TWO—BUILDING THE CONCEPTUAL MODEL

Operations Research & Synthesis is now in a position to define the situation *in such a form* as to allow logical and cogent development of the range of practicable alternative solutions. *The principal tool to this end is the logical structure or conceptual model which is considered by many to be the operating and distinguishing feature of Operations Research & Synthesis.*

The conceptual model is, in its essence, a presentation of the *relationships* that underlie the pertinent facts. Its purpose is to provide understanding how and why the facts behave as they do. In many cases the model makes it possible to make rational anticipations. It thereby makes possible purposeful action to adapt to the situation or to modify the relationships to obtain desired results. Such a model may be mathematical or non-mathematical as circumstances require. It is always the logical structure discussed in the preceding section.

Being a presentation of the relationships, every model is of necessity limited in its comprehensiveness and in its scope. It may be more or less accurate because the facts do not permit more

than approximation, or because of the objectives and of the assumptions of responsible managers. The model should, therefore, never be taken for more than a symbolic approximation to the real situation. It is one which rests on a human decision as to what factors in the situation are to be considered relevant. By definition, a model is an image—a symbol—or an analogy—and not an aerial photograph or miniature of the real situation.

For different purposes different models of the same situation may therefore be needed. A conceptual model may be primarily intended to display facts in a more significant manner, or it may be devised to provide a method for arriving at specific kinds of decisions logically and consistently. It may be a model of a very limited portion of an operation, or it may be a set of interlocking models that in effect describe an entire operation or an entire business. In the latter case, it is presumed that information can be bled off at various places so as to form the basis for making decisions—for instance, at various levels within the operation described—that are consistent with the objectives and purposes of the whole.

Actually, managers have long been working with such conceptual models—some indeed highly developed and highly useful over broad parts of a manager's work. An impressive and familiar example of an excellent conceptual model is the general accounting system.

It is designed to demonstrate the performance of an operation in terms of money. It establishes categories of inputs that may be measured. It establishes categories of outputs also measurable. In its way it demonstrates periodically the flow of the process through the operation in terms of dollars by stating values of inventory, receivables, accumulated costs and other related factors.

It is based upon certain assumptions, namely, that the dollar is a rational and stable unit of measure of the business operation, that matching current billing against current costs gives a reasonable measure of the profitability and soundness of an operation for most current purposes, that monthly and yearly measurement periods are the natural cyclic intervals in which to measure performance, that this type of general model applies to practically all types of businesses with relatively little modification.

It is a model from which alternative decisions may be derived. Of this there are abundant examples. It is also a model on which limited predictions may be based, particularly if the periodic readings from the model are adapted to purposefully establishing optimum trends and situations.

All in all, it is an excellent rational model of a business operation, sufficiently valid to meet the test of practicality for a wide range of managerial and financial situations. It is excellent also in the respect that it encompasses a view of the *whole* operation.

Despite its excellences, it is limited, too, since it does not describe, for example, the principles of flow or process applying to the manufacturing operation in such a manner that they can be used to

govern the flow or process, because of its orientation toward financial objectives and because it is, by design, a system of measurement based upon units of money and arbitrarily selected periods of time.

The accounting system, however, has, in effect, all of the characteristics of a conceptual model of the kind here being described. Among the families of models, it may be classified as a mathematical-digital-exact model.

A variety of classes of rational models are known to exist. . . . Indeed, the range of models is very wide. The simplest model may bring clarity through a graphical presentation which shows relationships between several sets of facts, such as relative rates of change, otherwise not apparent. Or a model may be needed requiring a set of equations to be solved simultaneously and into which a variety of values may be fed in order to survey the effect of variability between pre-set limits.

STEP THREE—USING MATHEMATICS

The conceptual model frequently has to be reduced to a mathematical formulation for the following reasons:

1. The nature of the available data and the inability to evaluate and understand them by any other method,
2. The complexity of the model and the consequent inability to study the effect of varying the factors involved in it through any less formal methods.

The use of mathematics in constructing the model or as the structure of the model itself is, consequently, a matter of economy rather than an inherent necessity.

STEP FOUR—TESTING OF ASSUMPTIONS

The validity of conceptual models for use in analyzing general, and often functional, managerial situations cannot usually be tested by the methods of controlled experiment, they have normally to be proved by their workability. Tests of this nature include:

1. Ability to describe correctly, and more clearly, known facts and situations,
2. Ability to describe reasons for situations that have existed in the past on the basis of the relationships among the principal factors of the model, that is ability to name causes of known effects,
3. Ability to carry the general relationships described back to a specific element and still observe the indicated relationship. For example, if the general principle states that all models of product of a given out-put, rated 10,000 volts and less, have the same magnetic structure, then each specific model examined, having this general description, should use the identical magnetic structure without exception regardless of whether rated 2,000 volts or 8,000 volts,
4. Varying the values of the principal factors in the model to test the consistency of the answers. This indicates the limitations of the model,
5. Variation of the principal factors in the model to test the plausibility of the answers. This indicates the validity of the model.

It is not until tests of this character have been made that the conceptual model may be considered a sufficiently

valid representation of the situation for practical application. It is therefore important to feed back acquired knowledge in order to modify either the assumptions on which the model is based or the operational structure of the model itself so as to make it fit more nearly the tests of practicality.

STEP FIVE—MAKING THE MODEL UNDERSTOOD BY OTHERS

One of the benefits of the well-conceived model is that it gives the full range of alternative decisions rather than a single answer. It will state the factors that are needed to effect the result and thereby provide the manager with a knowledge of what can be changed, and how much, to purposefully effect the result. In other words, it provides the manager with a logical framework in which to make a decision. The manner in which the factors are varied will depend upon the manager's evaluation of the risks, opportunities, costs, and practicalities of changing certain factors. Subsequently, changes in the factors and their confirmation under certain alternative assumptions will bring out the risk, opportunity, cost, and practicality of alternative courses of action.

As an oversimplified example, assume it is desired to increase return on investment. The accounting model discussed earlier suggests several alternative courses of action. These include increasing profits or reducing variable investment. Within limits either will accomplish the same results. But, since the accounting model in itself cannot give the whole picture required, it needs to be interlocked as required with other models based upon different concepts and units of measure. For example, when it is considered too risky to increase prices or cut costs, it may be elected to reduce inventory. The question then is specifically what portions of inventory should be reduced and how much in order that the ability to supply customers may not be adversely affected. For this consideration a model showing the quantitative relationships of inventories to finished product is required to provide the manager with further guidance—how far he may go in this direction without impairing the results. There are, therefore, two reasons why the manager must understand the model, its characteristics and its limitations, before he can take action on the basis of the results produced by the model.

In the first place, the model is no substitute for the manager's own judgment. It only makes possible better judgment and more effective judgment by showing the alternatives of action, the assumptions underlying each, the consequences of each, and the impact of any decision on all other phases of the operation. To make a sound and meaningful decision between alternatives, a manager must understand what he is choosing from. That means he must understand the model.

Secondly, no model comprehends the universe. It is a presentation of a small and limited aspect on the basis of definite assumptions. What the model can be expected to do and what it cannot be expected to do, what results it can be relied upon for and what results it cannot be relied upon for, its structural character (is it a radio or television

set?) and its limitations (is it a short wave or an AM receiver?) must all be understood to make possible sound judgment among alternatives, and to effectuate correct action for the decision taken.

STEP SIX—THE CLASSIFICATION OF ACTION ALTERNATIVES

Because of the inherent characteristics of models, the last stage in the research phase consists of acquainting the manager responsible for the final decision with the following:

1. *The characteristics and limitations of the model;* the kind of information to be obtained from it; the kind of input data it requires; and the kind of operations through which the given model produces meaningful alternative answers.
2. *The alternative decisions* derived from the analysis of the situation by means of the conceptual model; their risks, opportunities, costs, and practicalities and the impact each alternative decision would have on behavior of factors throughout the organization including the behavior of people required to make the decision effective.
3. *The kind of communication system* required to implement, measure, and govern action on all levels of the organization and to feed back to the responsible manager information regarding the effect of the action and to indicate to him when, as a result of changes in the situation, changes either in the details of the decision or in the basic decision itself are needed.

To this end, the action alternatives have to be classified according to their major characteristics. *Action* can be *classified* according to four criteria. Since all combinations of these criteria are possible, there are 12 classifications.

First criterion of classification: An action can be single, that is, an isolated decision requiring no action outside and beyond its own level or area on which it is taken. Or an action can be multiple, that is, it may require consistent action beyond its own level or area; or it may be itself affected by action taken outside its own level or area.

Second criterion of classification: A decision may be either non-recurrent or recurrent. A continuous action is a special case of recurrence and requires the same treatment.

Third criterion of classification: An action may be self-effectuating. The decision to separate a man from employment is, for instance, self-effectuating. (It is also a single decision.) A decision to curtail all appropriation requests by an arbitrary figure of 40% is self-effectuating. (It is also a multiple decision.) Or an action may be interlocking, that is, require additional balanced and consistent decisions either by the manager himself or beyond his level or area to become effectual.

Fourth criterion of classification: An action may be easily reversible. For instance, a decision to buy copper requirements against production schedule rather than on the basis of speculative guesses regarding the copper market. Or an action may be difficult to reverse —either because of the cost of reversibility or the time required. In extreme cases an action may be irreversible.

With the classifications of the actions,

the Research phase is concluded and the Action phase begins. The manager should now have available to him the following:

1. A description of the situation,
2. A presentation of alternative courses of action,
3. A description of the impact of each alternative course of action, its risks, opportunities and its impact upon the operations in general, and
4. The assumptions underlying the presentation of the situation, the scope and limitations of the presentation and of each alternative course of action.

C. ACTION PHASE

Taking action, the final phase—consists of *two main parts.*

One part is the making of the manager's decision, the description of which does not belong in this paper. It is reemphasized that Operations Research & Synthesis is no substitute for decision-making. It only facilitates understanding and so makes possible the taking of decision with fuller awareness of the nature of the situation, the range of alternatives to choose from, and their impact. *The decision itself must, of necessity, remain the manager's own, must be based upon his judgment and must be an exercise of his managerial responsibility.* Operations Research & Synthesis shows how to arrive at this decision with greater assurance and also points the way to the things which must be done to bring about its implementation more precisely.

The second part of the Action phase, however, brings use for Operations Research & Synthesis in again (after the manager's actual action decision has been made), in the design and installation of the Communications System through which the action is being made effective, through which its impact is being measured, and through which its results are being fed back to indicate the need for adjustment or for a new decision.

Except for an action of the simplest kind—that is, one that is single, nonrecurrent, self-effectuating, and irreversible—actions require such a Communications System specifically because:

1. Actions are *time-conditioned* in that they are based upon certain assumptions regarding the condition as of a given moment. They require adjustment or change if these conditions change.
2. Actions have a *time dimension* in that the action taken may not show effect until a certain time has passed.
3. An action usually *does not pay off* unless it has been in effect for a certain length of time.
4. All but actions of the simplest type require a sequential chain of consistent and balanced complementary action *by other people* with respect to the area directly embraced within the specific action and also with respect to other areas.

Because of these characteristics of action, the *model itself* may have to become *part of the system* through which the *action is being carried out.* In the first place, the model may be the way to organize for effective complementary action by others. This would be the case if in the judgment of the

manager the decision establishes a basic pattern for the handling of recurrent events. The model would then be used to delineate the range within which the actions of others should be expected to fall. The second use of the model as part of the system may be as a measurement for the effectiveness of action. The model may finally be used to measure the validity of the action and to indicate the need for changes.

These changes may be twofold. Adjustments may be required which, while leaving intact the basic assumptions and the model itself, modify the operations through which the action is being carried out. Or the actual result may fail to come up to the expected results—or the expected may not happen when it should happen—in which case the manager should be informed through the operations of the system of the need for a change in the decision or in the model underlying it.

In other words, the model can be built into the system to provide feedback, and thereby continuing steering and direction of the system itself.

. . .

The work here described is concerned with the acts, processes, and effect of operating. In this respect the word "Operations" in its title is descriptive. It defines the area of interest with respect to which it is used as being concerned with operations. *It is not concerned with the Work of the Manager as such, or with the managing of specific operations, but with the defining and description of processes of operation through which managers may most efficiently and expeditiously attain their chosen or defined objectives.*

Such result is attained not through Analysis alone, which is in effect the separation of a thing into its constituent parts. It is not obtained through Research alone which is critical and exhaustive investigation or experimentation having as its aim the development of new or the revision of accepted conclusions in light of newly discovered fact. It is attained by putting together those elements discerned by analysis, described and defined by research, so as to form an understandable and workable description of the whole operation. This is a Synthesis.

Hence, this work is most accurately and descriptively called "OPERATIONS RESEARCH & SYNTHESIS."

IN SUMMARY

Operations Research & Synthesis is a general process for constructing methodologies.

The process in three phases:

A Judgment Phase
A Research & Synthesis Phase
An Action Phase

One distinguishing feature of Operations Research & Synthesis is the use of a conceptual model which may be mathematical or non-mathematical. The general accounting system is one example of such a conceptual model.

Operations Research & Synthesis presents information and relationships to the manager that will assist in making and in implementing decisions.

Operations Research & Synthesis is not concerned with performing the

work of a manager as such, or with managing a specific operation. It is concerned with the defining and description of processes of operation through which managers may most effectively and expeditiously attain their chosen objectives. The decision itself, however, must of necessity remain the manager's own, must be based upon his judgment, and must be an exercise of managerial responsibility.

THE SCOPE AND LIMITATIONS OF OPERATIONS RESEARCH & SYNTHESIS

A. ITS SCOPE

Operations Research & Synthesis might be called a system of *qualitative logic* because of the following eight factors:

1. It finds and brings out the underlying patterns in the behavior of the business and in its environment, including those that have hitherto lain beyond the manager's field of vision or range of imagination.
2. It shows which factors are relevant (that is, are facts) and which are irrelevant (that is, mere data).
3. It shows the degree of reliability of the available data and what additional data are required to arrive at sound judgment.
4. It shows what resources will be needed in any of the alternate courses of action, and what contribution from each component or function would be required. It thus establishes clearly the relationship between ends and means.
5. It shows the limitations of each available course of action, its risks and its probabilities.
6. It shows what impact a given course of action would have on other areas, components and functions, the relationship between input and output; and the location and nature of bottlenecks.
7. It ties together the work and contribution of each function or component with those of all others, and shows their total impact on the behavior and results of the entire business.
8. Operations Research & Synthesis is a *method through which all managers on all levels and in all components can obtain information:*
   a. That is focused on the needs of the *business as a whole.*
   b. That, because based on mathematical and logical analysis rather than simply on the record of historical events, raises questions as to what underlies the phenomena rather than merely describing them, and hence focuses on action-creating decisions.
   c. That shows why a decision is needed and where, what requirements it has to satisfy, and what alternatives exist from which to choose.
   d. That defines the area of rational judgment based on currently available and usable facts.
   e. That makes possible decisions with a markedly higher degree of rationality in respect to their futurity, their risks, their probability and possibility for proper and concerted implementation at all

levels based upon completeness of knowledge and understanding.

This is clearly the kind of information a manager—whether a general or a functional manager—needs to do his own work in such a manner as to contribute the most to the attainment of the objectives of the business and to its over-all results.

B. THE PROPER USE OF OPERATIONS RESEARCH & SYNTHESIS

Because Operations Research & Synthesis is so potent a method, it has to be used properly; otherwise its use might cause harm as well as cause good.

The most productive application of Operations Research & Synthesis is, as has been said before, for the purpose of finding, presenting, and making possible, where needed, modification of the characteristics of a business as an organic and dynamic entity.

Conversely, Operations Research & Synthesis can conceivably do real harm if used to "solve" problems in one area and function as if they were independent and existing by themselves. This would make almost inevitable "sub-optimization," that is a solution which achieves the optimum for one area or function, but at the expense of another area or function and thereby of the whole business.

The best way to use Operations Research & Synthesis would therefore seem to be to use it first to establish a foundation in the form of a definition of the characteristics of an integrated business. This definition can be quite wide, but applications of Operations Research & Synthesis to individual areas and functions can be very refined, yet escape the danger of sub-optimization if capable of being related to such a common foundation representing the integrated business.

C. WHAT OPERATIONS RESEARCH & SYNTHESIS IS NOT

In introducing a new method, its limitations should always be as carefully indicated as its potential. It is, therefore, important to spell out first what Operations Research & Synthesis is not:

1. *Only the Tools are New*—The basic method itself is not "new"; only the tools are. What Operations Research & Synthesis enables the manager to do has been done for a long time by a few people with exceptionally perceptive imagination. But Operations Research & Synthesis enables *every* manager to do well and systematically what hitherto only the rare "natural" could do. It converts an "art" into a scientific discipline that can be taught and learned. It substitutes method and principles for hunch.
2. *No Substitute for Judgment*—The manager will always have to exercise judgment, will have to make decisions and state objectives. But, Operations Research & Synthesis gives him the kind of factual information that makes better and sounder judgment possible by establishing clearly the character of the decision, the range within which judgment must be exercised, and the fundamental factors and variables to be taken into account in both making and implementing the decision or objective.

3. *Not an Additional "fifth element" of the work of a Manager*—Operations Research & Synthesis is not a new element in the "Work of a Professional Manager," nor a substitute for any of the four basic elements therein. It is a method to make simpler and more effective the Work of the Professional Manager in all four areas—Planning, Organizing, Integrating, and Measuring—by giving the manager improved tools of definition and measurement.
4. *No Substitute for Sound Organization*—Operations Research & Synthesis is not a means to obtain effective performance despite unsound organization of component and work. It does not eliminate the need for clear organization structure and relationships, for clear definition of responsibility, or for proper delegation of decision-making authority to the managers, and individual workers, in all functional and sub-functional areas. On the contrary, it both requires sound organization and strengthens it, by eliminating need or justification for such organizational makeshifts as "coordinators" or decision-making committees.

D. ITS LIMITATIONS

Finally, Operations Research & Synthesis is not a universal method. In particular, two of the most important tasks of the manager are not susceptible to attack or solution by the methods of Operations Research & Synthesis.

In the first place, the manager of a business, unlike the natural scientist, has to take action and often on a *time* basis not fully under his control. This means that he cannot confine himself to understanding the situation or to adapting his behavior to outside forces. He has both the responsibility and the opportunity to modify the outside environment, if not to create it. Operations Research & Synthesis can help the manager in this task by identifying the factors that will have to be modified to take possible successful action toward the desired goal. It can identify alternate courses of action. It cannot, however, tell him when to take such action or what action to take.

Secondly, the manager, unlike the natural scientist, does not deal with inanimate nature but with human beings. Human beings, however, cannot be quantified. Decisions affecting human beings, therefore, always require moral and ethical principles and so are by definition fundamentally not the kind of facts within the scope of Operations Research & Synthesis. It should be kept in mind, however, that Operations Research & Synthesis *may* indicate the extent to which consideration for human beings becomes a factor in decision and action.

IN SUMMARY

1. Operations Research & Synthesis is a conceptual approach to the definition of business situations and a systematic method to increase understanding and so to facilitate rational business decisions and to help make them fully effective. It is not a bag of tricks.
2. Operations Research & Synthesis provides comprehensive methods of analysis and synthesis using *both* mathematical and non-mathematical

tools as the situation requires. It is not just the application of mathematical techniques to business problems.

3. While not unlimited, Operations Research & Synthesis applies to a wide variety of business decisions ranging from the determination of the optimum speed of a machine feed or the best size of a salesman's territory to the finding, presentation and modification of the basic characteristics of a business. *Its most important contribution may well be that it helps a manager to see and understand a business, or a situation, as one integrated whole* despite today's technological and economic complexities, the variety of functions each of which has its own highly developed and specialized skills, and the overlapping and cross-cutting of divergent time dimensions for different processes or operations. It may be described as the process of designing to meet the described situation either through adaptation to or modification of the situation.
4. Finally, Operations Research & Synthesis is an aid to the Professional Manager. It is no substitute for the making of risk-taking decisions by a manager, nor for any of the four basic elements of his work, namely, exercising leadership through planning, organizing, integrating, and measuring. It is not a "mathematical brain" any more than the computer is an "electronic brain." On the contrary, it is one more, and helpful, means to make the manager more capable of being a truly Professional Manager.

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE DEVELOPMENT OF OPERATIONS RESEARCH AS A SCIENCE

RUSSELL L. ACKOFF

Ten years ago it would have been difficult to get an operations researcher to describe a procedure for conducting OR. Today it is hard to keep one from doing it. Each practitioner's version of the operations-research method (if recorded) would differ in some respects. But there would also be a good deal in common. For example, I think most operations researchers would agree on what are the major phases of an OR project. Let me expose myself by suggesting the following phases of an OR project as representing a common agreement:

1. *Formulating the problem.* This refers to both the consumer's [decision-maker's] problem and the researcher's problem.
2. *Constructing a mathematical model to represent the system under study.* This model expresses the effectiveness of the system as a function of a set of variables

Operations Research, *June 1956, 4:3, 265–287.*

at least one of which is subject to control. Variables of either type may be subject to random fluctuations, and one or more may be under the control of a competitor or other 'enemy.'

3. *Deriving a solution from the model.* This involves finding the values of the 'control variables' that maximize the system's effectiveness.

4. *Testing the model and the solution derived from it.* This involves evaluating the variables, checking the model's predictions against reality, and comparing actual and forecasted results.

5. *Establishing controls over the solution.* This involves developing tools for determining when significant changes occur in the variables and functions on which the solution depends, and determining how to modify the solution in light of such changes.

6. *Putting the solution to work.* Implementation.

It is not implied that the steps enumerated are ever conducted in this order, or that one step must be completed before another is begun. In many projects, for example, the formulation of the problem is not completed until the project itself is virtually completed. There is usually a continuous interplay between these steps during the research; that is, there is usually considerable recycling of the results of each step through the preceding steps. . . .

FORMULATING THE PROBLEM

It is useful to distinguish between the consumer's (decision-maker's) problem and the research problem, even though they are closely related. The latter is a transformation of the former, primarily involving the definition of a scientific basis for selecting a course of action as a "solution."

THE CONSUMER'S PROBLEM

The consumer's problem is seldom given to the operations-research team; it is extracted by the team from reported symptoms and the analysis of the system involved. Formulation of the consumer's problem usually requires the following steps:

1. Identification of those in control of the operations under study and analysis of their decision-making procedure.

2. Determination of the decision-maker's objectives. These fall into two classes: those to be obtained and those to be retained. The former constitute goals; the latter provide the restrictions on the problem.

3. Identification of other participants: those who carry out the decisions and those who are affected by them, including 'enemies.'

4. Determination of objectives of other participants which can affect their responses to decisions in the area under study.

5. Analysis of the processes or operations whose control is involved in the decision area under study.

6. Determination of alternative courses of action available to the decision-maker.

7. Determination of the alternative courses of action available to the other participants, action which can affect the outcome in the area under study.

It is apparent that in the process of formulating the consumer's problem the operations-research team analyzes the system under control and the organization and procedure by which it is controlled. Consequently, OR has come increasingly to realize that the types of

systems it studies involve *organized human behavior* as well as physical objects and their behavior. Little wonder, then, that OR is turning more and more of its attention to the work of others in the area of organizational behavior, and that OR has begun to work more and more in this area itself. . . .

THE RESEARCH PROBLEM

In most general terms, the OR team's problem is to determine which alternative course of action is *most effective* (optimum) relative to the decision maker's set of pertinent objectives. Consequently, in formulating its problem it must define the measure of effectiveness to be used, and the meaning of 'most' in 'most effective.' The steps involved may be enumerated as follows:

1. Definition of the measure of efficiency to be used relative to each objective to be obtained (goals).
2. If the units in the measures defined are not commensurate, selection of a common measure (standard) of efficiency and transformation of the measures obtained into the common measure by either (*a*) finding an objective transformation (e.g., finding the dollar value of waiting time or good will), or (*b*) finding a subjective transformation (e.g., determining the relative importance or utility of the objectives to the decision-makers).
3. Definition of 'most effective' (e.g., maximum expected profit, minimum expected waiting time, etc.). This, in effect, defines a 'best' or 'optimum' decision.

A great deal of study has been directed to defining 'best decisions,' particularly since the pioneering work of mathematical statisticians (such as Wald), of mathematicians (such as von Neumann), of economists (such as Arrow), and of philosophers of science (such as Dewey and Singer). This area of inquiry was recently named *Decision Theory*. . . .

The main effect of this development on the practice of OR has been the growing realization that there are decision objectives other than maximizing expected return and minimizing maximum loss. That is, in many practical situations there are criteria of optimality that are more appropriate than these two mentioned.

CONSTRUCTING AND SOLVING MODELS

Operations research has reached a point in its development at which an OR model can be defined as a mathematical representation of the system under study, a representation which takes the form $E = f(x_i, y_j)$, where E represents the effectiveness of the system; x_i represents the variables of the system that are subject to control; and y_j represents those variables not subject to control. The restrictions on values of the variables may be expressed in a supplementary set of equations and inequations.

Over the history of OR, certain types of processes or systems have been encountered repeatedly. The structure of these recurrent processes has been abstracted and analyzed so as to yield what we might think of as prototype models. Though these prototype models can seldom be applied in a specific situation without adjustment, they provide a valuable point of departure. These are the tools with which OR is

beginning to fill its kit.

Recognition of recurrent processes has also led to abstraction and definition of these processes and the problems emerging from them. Thus the body of concepts in OR has been growing in size and precision. As yet, a completely common vocabulary has not been reached, but we do have the beginnings of an OR terminology.

. . .

Seven types of processes or systems have been identified: *Inventory, Allocation, Waiting-line, Routing, Replacement or Renewal, Information collection, and Competitive.*

The most conspicuous and copious developments in OR are the models applicable to these processes. . . .

DERIVING SOLUTIONS FROM MODELS

Before concluding this discussion of models and their solutions, several words should be said about some very general developments in this area.

First, note should be made of the use of operational experiments and operational gaming. These are methods of simulating pertinent aspects of a system under controlled conditions so as to obtain some estimates of the system's properties. These estimates can then be used in the model of the system. Operational experiments on waiting-line processes, for example, have been conducted at the Massachusetts Institute of Technology in which electronic devices simulate random arrivals. Operational gaming, which is experimentation involving people as decision makers has been going on at Rand, the Operations Research Office, and Tufts College. Case Institute of Technology has used operational gaming in an industrial project to determine how effectively production planners could use certain planning aids developed during the project. Alderson and Halbert have used the method extensively in their studies of consumers' buying habits. This unique application has yielded quantitative data and insights into a very complex type of behavior. . . .

Second, note should be made of the advance in the art of using Monte Carlo procedures for deriving solutions from models. . . . The art of using computers in Monte Carlo runs is now a highly developed one. Finally, note should be made of the value of Rand's publication of extensive random and random normal numbers.

TESTING THE MODEL AND SOLUTION

A model is never more than a partial representation of reality. It is a good model if, despite its incompleteness, it can predict the effect of changes in the system on the system's over-all effectiveness to acceptable and useful accuracy. The adequacy of the model can be tested by determining how well it does predict the effects of these changes. These predictions can be checked either prospectively or retrospectively. The latter is generally more feasible since it minimizes disruptions in the operation of the system, but it requires good data on past operations. Prospective tests are usually done on a small-scale or trial-run basis.

Operations research, in general, has

become increasingly sophisticated in the use of statistical methods for testing its models and solutions. For example, the use of 'designed experiments' (in the Fisherian sense), the analysis of variance, and the analysis of covariance has become more widespread. These methods are used to test the significance of the contribution of variables to the system's effectiveness. In addition, such standard statistical tools as the t-test, the F-test, and the x^2-test are being used more frequently as well. This turning toward statistics reflects, in part, the increased participation of competent mathematical statisticians in operations research.

CONTROLLING THE SOLUTION

A solution derived from a model remains a solution only as long as the uncontrolled variables retain their values. The solution itself goes 'out of control' when the value of one or more of these variables has changed significantly. The significance of the change depends on the amount by which the solution is made to deviate from the true optimum under the changed conditions and the cost of changing the solution in operation.

Designing a control procedure for a solution is not yet widespread in OR. Since more and more problems studied by OR involve the development of an optimal decision rule to be used repetitively over a long period of time, this aspect of its methodology should receive increasingly more attention in the next few years.

To set up complete controls for a solution, the following steps are required: (a) For each variable and relationship which appears in the model, define a significant change. (b) Set up a procedure for detecting the occurrence of such significant changes. (c) Specify how the solution should be modified if such changes occur. Operations research is using the tools of statistical quality control to detect the kind of changes involved here. It has come to learn that these tools provide ready-made devices for use in this control function. More widespread acquaintance with them is to be expected.

IMPLEMENTATION

Proper implementation of a solution to a problem is perhaps still more an art than a science. It is true, however, that operations research is learning a good deal about the kinds of people and organizations with which it deals. This knowledge is finding its way into guiding principles which operations-research analysts discuss almost every time they gather.

One of the things OR has learned about putting results to work is having a considerable effect on its methods. Solutions are generally carried out by personnel whose mathematical sophistication is less than is desirable. Consequently, if the OR team wants to assure use of its recommended decision-rules, it must simplify the rules handed over to executives and operating personnel. In many cases this means the team must either translate elegant solutions into approximations that are easy to use or sidestep the elegance and move directly to a quick-and-dirty decision-rule. Operations research is learn-

ing that an approximation that is used may be a great deal better than an exact solution that is not.

In some problems the urgency attached to obtaining a solution, or the limitation of resources, may also require direct movement to quick-and-dirty solutions. In many cases such 'solutions' need not deviate from the optimum (in the purist's sense) by too much. It is OR's job to see that as little sacrifice of effectiveness is made as is possible.

Operations research has learned another general lesson from its involvement with implementation of results. A solution must be 'spelled out' in the language of those who will use it. In the process of translation and operation the OR team almost always finds aspects of the situation which it had not taken into account, aspects for which adjustments in the proposed solution are usually required.

There is a long distance between a recommendation and a successful application. Operations research is developing a healthy respect for the difficulty of covering that distance.

part ✦ 2

The Methodology of Operations Research: Models and Model Building

OR, as you have seen, applies the scientific method to the study of business problems. It applies, in particular, the method of research which has been used with such great success in the natural sciences. An important ingredient in this method is the formulation of an hypothesis or theory regarding the nature of the mechanism underlying a phenomenon. This theory is then tested against observed facts and modified in the light of the test results. The modified theory is then tested and itself modified and this process is continued until the scientist is satisfied that his theory accounts for the observed facts with sufficient accuracy for his purposes.

Scientists can rarely obtain interesting and useful results by studying a phenomenon directly as it takes place in nature. More commonly, in order to test a theory, they must reproduce the mechanism that they believe to be responsible for the phenomenon, under controlled conditions in a laboratory. They must construct a replica of the cause and effect relationships which, as they see it, are producing the phenomenon and must study and analyze this replica, rather than the natural phenomenon in its natural setting. The replica is, therefore, an embodiment in physical, graphical or mathematical form of the scientist's theory of

the origin or nature of a phenomenon. It is a representation of the underlying causal mechanism. It is not, however, a perfect and complete representation but contains only those elements which the scientist considers important.

In the language of science such replicas are known as *models.* And in the methodology of science, models occupy a key position. It follows that models are of central importance in the methodology of OR. For this reason, the second section of this book is devoted to a discussion of the nature of models and their role both in scientific work generally and in OR particularly.

The first selection, authored by an eminent medical statistician, is an excerpt from a notable book which successfully translates the basic ideas underlying the use of statistical tools for decision making from the language of mathematics to plain English. It discusses the nature of models, their types, their advantages and disadvantages and their role in scientific thinking.

In the second selection, a prominent social scientist offers another view of the nature of models and their general function in attempts to understand a structure or process. In addition, he considers some yardsticks by which the performance of models may be evaluated and discusses some possible misuses of mathematical models.

The next three selections appeared originally in the monthly publication of a marketing and management consulting firm, Alderson Associates, Inc., which has been a leader in applying OR to marketing problems. In these selections, some of the problems encountered in developing and using mathematical models are explored and the procedure involved in selecting an appropriate model for a problem is described.

A widely read book devoted principally to one of the more important techniques of OR—mathematical programming—is the source of the sixth selection. Its author, an early contributor to the development of OR in business applications, enumerates succinctly the advantages to be gained from the use of mathematical models.

Since mathematical models occupy a key position in the methodology of OR, a close examination of their constituent elements as well as of the procedure followed in their construction is warranted. Such examination, at the microscopic level, may contribute to a fuller understanding of the nature and utility of mathematical models. The seventh and eighth articles in this section provide this close examination.

Authored by the Manager of Market Research of the International Business Machines Corporation, the seventh selection is a discussion of the kinds of equations used in the construction of mathematical models. In addition, it illustrates the process of putting together a mathematical model for a marketing problem and notes some common arguments for and against the use of such models.

The process of constructing a mathematical model is also the subject of the eighth article. Written by the same man who wrote the sixth selection, this article describes in detail the procedure followed in translating a problem into its mathe-

matical analogue. This is done within the context of a production rather than a marketing problem.

The final selection in this section is from an article written by the Director of the OR Group at the Case Institute of Technology. It describes the types of models that are most often used by operations analysts and identifies the techniques with which each of these models is associated.

✦✦✦✦✦✦✦✦✦✦✦✦✦ MODELS

IRWIN D. J. BROSS

THE SYMBOLIC WORLD

. . . I want to devote some attention to the broad concept of a *model.* Models are vitally important in scientific work and, in my opinion, in any intellectual endeavor. An understanding of the nature and role of a model is prerequisite to clear thinking.

In ordinary language the word "model" is used in various ways. It covers such diverse subjects as the dolls with which little girls play and also the photogenic "dolls" who occupy the attention of mature men. I shall be concerned here with model in the sense of replica (as in a model airplane).

PHYSICAL MODELS

There are several kinds of model aircraft. Solid scale models resemble the actual planes in general appearance (shape, markings, etc.). The flying model aircraft not only resemble the originals in appearance but, to some extent, in *function* as well (i.e., they are capable of free flight). Some very elaborate models are essentially simplified versions of real aircraft; they have gasoline engines, operable controls, and may even have radio-control mechanisms which allow the plane to be directed from the ground.

A boy who is interested in aviation can learn about the subject from the construction and operation of such flying models. In much the same way a scientist who has constructed a model of some natural phenomenon may learn about this phenomenon from a study of his model.

The model aircraft is easier to study than a full-sized aircraft for various reasons. It is more convenient to handle and manipulate. It is also simpler than the original, and principles of operation may be more apparent. There is some danger of over-simplification, of course, and some characteristics of a real aircraft would be overlooked if all attention were focused on the model.

As a matter of fact, adult scientists

use model aircraft to learn about the performance of full-sized aircraft. They build carefully scaled replicas and test these models in wind tunnels. This is a much more economical process than to build a full-sized airplane and then to test *it* in a wind tunnel (a mammoth wind tunnel is a fabulously expensive piece of equipment). This type of argument by analogy has proved quite successful and is used all the time by aircraft engineers.

I do want to emphasize that the aircraft engineers do not trust the method entirely, that they carefully test the full-sized aircraft as well as the model. In other words, it does not follow that one can *automatically* obtain useful information about the original phenomena from the study of a model. Whether a model will be useful or not will have to be learned from experience, by comparing the performances of the original phenomenon and the replica.

The model represents a process of abstraction. The real aircraft has many properties or attributes such as shape, weight, and so on. Only a few of these properties are duplicated in the model. The wind tunnel model, for example, duplicates only the shape. However, the aerodynamic performance depends largely on this one characteristic; the other properties are more or less irrelevant.

This is an example of an effective process of abstraction. It allows us to focus our attention on a much simpler phenomenon without much loss from the fact that many details have been neglected.

This particular type of abstraction, the construction of a physical model, is used in various branches of science, engineering, and industry. Models are used to design ocean liners, bridges, water supply systems, and all sorts of products from automobiles to stage scenery. Not all models involve a change in size. In aircraft construction, for example, a full-sized model of a part of a plane is sometimes constructed out of wood in order to insure that an absent-minded designer does not put components in places which cannot be reached for repairs. In this situation the relevant factor is size, and the mock-up (as it is commonly called) eliminates other factors such as weight, function, and so on.

ABSTRACT MODELS

In the scientific world physical models are occasionally used for instructional purposes. In a planetarium you will generally find a model—little spheres which revolve on wire arms around a big sphere—which presents a picture of the astronomer's conception of the solar system. This sort of model is often used to demonstrate a phenomenon such as an eclipse. A rather similar physical model is sometimes employed to explain the atom to the general public. The solar model and the atom model illustrate one striking and sometimes confusing characteristic of models; two very diverse phenomena can sometimes be represented by similar models.

The solar model which you can see in a planetarium has had a very interesting history. Nowadays we think of the sun as a giant globe with a large family of little spheres circling around

it. We locate ourselves on the third little sphere (counting out from the sun), and this notion does not cause us any mental anguish. In earlier days the picture was quite different and the earth was regarded as the center of the system. Of course if we go back still further there are all sorts of fabulous models which involve giants, turtles, and sea serpents. The history of astronomy is the story of the evolution of a model.

Did you notice that in describing the solar model I was actually taking a further step in abstraction? I was going from a physical model to a *verbal* model. The little balls were replaced by their symbols, the words "little balls."

All of us are accustomed to using verbal models in our thinking processes and we do it intuitively. Verbal models have played an important role in science, especially in the preliminary exploration of a topic and presentation of results. Verbal models are subject to a variety of difficulties, some of which I have discussed earlier, and most scientific fields have advanced (or are trying to advance) to the next stage—symbolic models of a mathematical nature. Astronomy was one of the first subjects to make this transition to the symbolic model. It should be noted that *until* this stage was reached there was really no reason to prefer a model with the sun as a center to a model with the earth as a center.

SYMBOLIC MODELS

In a symbolic model the balls and wire arms of the physical model of the solar system are replaced by mathematical concepts. Geometrical points are substituted for the balls. The next problem is to replace the wire arms which hold the balls in place. Now the wire arms have fixed lengths, and these lengths can be stated numerically. If all of the little balls revolve in the same plane, only one additional number is needed to locate the geometrical point. This number would be the angle between the wire arm and a stationary arm which would serve as a reference point.

Hence two numbers—the radius (length of arm) and an angle—will fix the location of the geometrical point just as effectively as the wire arm fixes the location of the little sphere in the physical model. Actually the astronomer's model is much more complicated than the symbolic model which I have described, but the general principle of construction is the same.

Now suppose that the astronomer wants to use his model to predict eclipses. He will have to take observations to obtain specific numbers to use for the radius and angle. These empirically determined quantities are substituted in the mathematical model and, after various manipulations, the astronomer announces: "There will be an eclipse of the moon visible in the northeastern part of North America on such-and-such a date and at so-and-so time."

It is at this point that a comparison of alternative models can be made. If the predictions are borne out, the successful model can be used for future predictions. If, on the other hand, the eclipse does not occur at the specified time, the scientist must begin looking for another model.

The Ptolemaic astronomers set up a

mathematical model of the solar system with the earth as a center. They first considered that the other astronomical bodies moved in circles. When this picture did not lead to adequate predictions the Ptolemaic astronomers decided the paths of the heavenly bodies were epicycles. If you would like to visualize an epicycle, imagine two gears, one large and standing still and the other small and rolling around the rim of the large one. An epicycle is the path of a tooth of the small gear.

This complication led to a little improvement in prediction, but the forecasts were still quite unsatisfactory so the model was complicated still further. This time the astronomers postulated that the paths of the heavenly bodies were epicycles *on* epicycles, literally a "gears within gears" situation.

If you think that this is getting too complicated consider the sad plight of the astronomers. *They* had to make the calculations which go along with this model of the solar system. Nonetheless it was many years before the simpler model with the sun at the center of the solar system was widely accepted.

There is a moral in this epicycle story. Scientists occasionally become attached to a model even though it does not give adequate prediction. They try to use the model by cutting off a piece here or adding a piece there. This patchwork can go on for many years, and the resulting crazy quilt may prevent the development of new and more efficient models. After all, when it takes a scientist ten years to master a complex model, he has a vested interest in it, and he sometimes is hostile to labor-saving devices which may deprive him of his job. "Epicyclitis" is a symptom of senility in a scientific field.

MATHEMATICAL MODELS

It might be puzzling to understand why the astronomers should go from a nice simple physical model with little spheres on wire arms to a symbolic model with all sorts of queer mathematical signs when, if sufficient care were taken in the construction of the physical model, it would be possible to use it directly in order to predict eclipses. The astronomer's choice is a matter of taste. From the astronomer's point of view it is the mathematical model which is the *simple* one and the physical model with balls and wire which is complex. Since the physical model is made out of metal it not only has attributes which are intended to simulate the solar system, but it also has a lot of attributes which depend on the materials used in its construction and the way in which it is made. Thus the wire arms can be geared to rotate at an appropriate speed but the mounting and drive arrangements of the model are attributes of the model and *not* attributes of the solar system which it is supposed to represent.

Even though great care is lavished on the construction of the physical model the predictions which would come out of it would depend on friction, vibration, and other characteristics of the *model.* Hence the prediction would be rendered inaccurate by the entrance of attributes other than the ones which were deliberately built into the model to simulate the solar system.

In a *mathematical* model, on the other hand, the material of the model itself—in this case the symbolic language—does not ordinarily contribute such extraneous and undesirable attributes. If we want friction in the mathematical model we can put it in symbolically, but otherwise this friction will not appear in the model and hence cannot disturb our predictions. In the physical model the process of abstraction tends to introduce new and irrelevant details, while in the mathematical model the process of abstraction does not.

In this sense, therefore, a mathematical model is simple whereas a physical model is complex. It may strike you as curious that I should say that Einstein is working with an extremely simple model in his theory of relativity, while a schoolboy is working with an extremely complex model when he builds an airplane. If you think it over carefully, however, you may see the justice of the statement.

Now and then a mathematical model gets beyond the resources of the mathematicians who construct it, so a physical model is substituted to obtain an answer. This is done in the Monte Carlo method, a device for solving mathematical problems by having one of the giant brain computers play gambling games with itself. However, such devices are used for computational convenience rather than conceptual simplicity.

The construction of symbolic models is an important part of the job of the scientist, and the great advances in science are those in which a useful new model is introduced. In physics the powerful model devised by Isaac Newton is one landmark, the relativity model of Einstein is another, and the quantum models are a third landmark. In chemistry the gas laws, the mass action laws, and the periodic table are all the end results of successful models of atomic and molecular processes. In biology the evolutionary model of Charles Darwin (a verbal model) has been developed into a mathematical model by R. A. Fisher and Sewell Wright. Another important biological model is the one which describes genetic inheritance. In medicine the models are mainly verbal, but they are of great importance. Harvey's model of the circulatory system, and the various models of the reaction of the human body to invading organisms have influenced the development of the modern treatment of diseases.

Effective verbal models which describe the transmission of disease have been useful in the eradication of many of the epidemic diseases which used to terrorize humanity. Efforts are currently in progress to translate these verbal models into mathematical ones (epidemic theory), but the earlier models have been so successful that a modern investigator is often hard put to find enough data to test his new mathematical models!

Currently, there is research under way which is attempting to devise mathematical models for sociological phenomena, such as the growth of cities, and for psychological phenomena. Norbert Wiener in *Cybernetics* [1] deals with the mathematical model as-

[1] Wiener, N., *Cybernetics,* John Wiley & Sons, Inc., New York, 1948.

sociated with the operation of the human brain.

One of the key steps in the progress of a field of knowledge toward scientific maturity is the fabrication of models which enable successful prediction in that field. A tremendous amount of imagination and insight is needed for the creation of new models, but they are only half of the story. The mere creation of models is not enough; the models must survive exacting tests, they must meet the pragmatic criterion, they must work.

This brings us back to data. The test of the model involves data from the real world. Without adequate data the construction of models is a mathematical pastime. Purely speculative mathematical models may be as useless as purely speculative verbal models. For example, I might construct a very fancy mathematical model to describe the mechanism of transmission of some virus disease. No good diagnostic test may be known for the disease, and consequently the available data may be quite unreliable. If a doctor comes along with a quick, cheap, and effective skin test for this disease, it may then be possible to get adequate data to test my fancy model. Until this happens my model is just another mathematical game. After the development of the skin test, the model may turn out to be useful in the understanding and control of the disease or, as is more likely, it may turn out to be a complete waste of time.

Progress in science is based on this constant interplay between model and data. Sometimes there is a tremendous amount of observational data available but no satisfactory model, so that little progress is made. This was the situation in astronomy before the heliocentric model and it also has occurred repeatedly in the biological sciences. At other times there are elaborate models but little adequate data. Something resembling this situation occurred in economics where an elaborate mathematical theory was developed which did rather poorly when tested with actual data.

Occasionally a scientist not only works out the model but also obtains the data. Darwin and Galileo accomplished this feat. More often one man, such as Brahé, gathers good data and another man, such as Kepler, supplies the model. When this division of labor occurs it is rather pointless to say that the model-maker is a greater scientist than the data-grubber, for the advance depends on teamwork.

ADVANTAGES

Why should a model be used? The real answer to this question is that this procedure has been followed in the development of the most successful predicting systems so far produced, the predicting systems used in science. It is simply a matter of going along with a winner.

Some of the advantages of model-making might, however, deserve a separate statement. A big advantage of a model is that it provides a frame of reference for consideration of the problem. This is often an advantage even if the preliminary model does not lead to successful prediction. The model may suggest informational gaps which are

not immediately apparent and consequently may suggest fruitful lines for action. When the model is tested the character of the failure may sometimes provide a clue to the deficiencies of the model. Some of the greatest scientific advances have been produced by *failure* of a model! Einstein's work was the outgrowth of the Michelson-Morley experiment in which the aether model led to unsuccessful prediction.

Another advantage of model-making is that it brings into the open the problem of abstraction. The real world is a very complex environment indeed. An ordinary apple, for example, has a great many properties—size, shape, color, chemical composition, taste, weight, ad infinitum. In making a decision about the apple, such as whether to eat it or not, only a few of these characteristics are considered. Some degree of abstraction is necessary for decision.

The model-maker must, therefore, decide which real world attributes will be incorporated in the model. He may decide that the size of the apple rather than shape is important to decision. He may, if he is setting up an inspection plan, concentrate on the number of worm holes. If he is interested in the velocity of a falling apple, on the other hand, he may include only the weight of the apple in his model.

By making this process of abstraction deliberate, the use of a model may bring such questions to light. Moreover, it may suggest preliminary experiments to determine which characteristics are relevant to the particular decision problem under consideration.

Once the problem is expressed in symbolic language there is the advantage of the manipulative facility of that language. The symbolic language also offers advantages in communication. It allows a concise statement of the problem which can be published. Moreover, it is more easily integrated with the other scientific work which is also in symbolic language.

Another advantage of mathematical models is that they often provide the *cheapest* way to accomplish prediction. Sometimes it is possible to reach the same results by the sheer mass of data —by a "brute force" attack on the problem—but the mathematical route is generally more economical.

One reason for this is that a newly-minted Ph.D. in mathematics can be hired (alas) for a salary which could not entice a good plumber. A Ph.D., a pencil, and some paper may be all the equipment necessary to handle the symbolic manipulations of the model. Only a very small proportion of the millions currently spent for research goes into model-making. Even when the scientists are well paid, most of the money goes into the process of collecting data.

DISADVANTAGES

The use of models also has some drawbacks. The model is subject to the usual dangers inherent in abstraction. A mathematically feasible model may require gross oversimplifications. There is no guarantee that an investment of time and effort in constructing the model will pay dividends in the form of satisfactory prediction. No process, however, can provide such a guarantee.

The symbolic language is also sub-

ject to limitations. It may be beyond the ability of a mathematician to manipulate the symbolic language so as to obtain useful results. In such cases it may be more efficient to use direct methods. In gambling-game problems, such as the game of solitaire, it may be easier to play a large number of solitaire games and determine the probabilities by the Direct System than to embark on a mathematical analysis of the probabilities.

There is another very grave danger in the use of models. After a scientist plays for a long time with a given model he may become attached to it, just as a child may become, in the course of time, very attached to a doll (which is also a model). A child may become so devoted to the doll that she insists that her doll is a real baby, and some scientists become so devoted to their model (especially if it is a brain child) that they will insist that this model *is* the real world.

The same sort of thing happens with verbal models, as the semanticists point out, when a word and its counterpart in the real world are regarded as the same thing. This identification in the world of words has led to unhappy results which are reflected in the real world. The behavior of individuals who are unable to distinguish between words and the real world may become so bizarre as to lead to the classification "insane."

Now things are not this bad at the scientific level largely because of the self-corrective features of the sequential process of model-making which provide a periodic return to the real world after each excursion into the symbolic world. The test of the model acknowledges, as it were, the supremacy of the real world. If the model fails to predict what will happen in the real world, it is the model that must give way. This is the standard of scientific sanity.

When this standard is not admitted, a conflict between a model's predictions and happenings in the real world will sometimes lead instead to the rejection of the real world. This course is the prelude to disaster. To guard against such disasters it is well to remember the following rule for working with models: A model is neither true nor false.

The standard for comparing models is utility, i.e., successful prediction. The evaluation of a model is therefore dependent on the situation in which it is to be used; it is not *intrinsic* (i.e., dependent only on the model itself). If this point is understood several apparent paradoxes in science disappear.

One such paradox is the simultaneous use of two contradictory models. An example of this paradox occurs in the field of physics in which a *wave* and a *photon model* for light are both accepted. Wave theories are used when *they* provide successful prediction, and in other situations the photon theory is employed. Hence the paradox arises only if the models are identified with the real world.

Another paradox is the occurrence of scientific revolutions which (unlike political revolutions) do not interrupt the orderly development of the area. If models are not identified with the real world, the revolution is merely the substitution of a refined model for a cruder earlier model. Most of the time the

older theory continues to be useful in the original applications; it is only in extended applications that the newer theory gives better prediction. The older theory is often a special case of the new theory. This explains why, despite the revolutionary work of Einstein, the older Newtonian physics is still used. In designing a dam or bridge, for example, both models would lead to essentially the same predictions (or in other words, the predictions are indistinguishable at the practical level).

One class of scientific workers does not worry about the testing of its models. They are the mathematicians. Their only interest (as long as they are functioning as mathematicians) lies in symbolic derivations from the models. Their business is to provide models in which the symbolic implications are worked out—anyone who wants to use the model for real world predictions will have to test it first. Nevertheless, the mathematicians serve a useful purpose in society (though a pure mathematician would strenuously deny it) by providing the scientists with ready-worked models. Often the models created by mathematicians are not used for years, or even centuries, but the literature of mathematics is a sort of Sears-Roebuck catalogue of models which may be consulted whenever a special type of model is needed. Unfortunately it takes some mathematical sophistication in order to use this catalogue.

As long as the model is completely divorced from the real world the criterion of utility cannot be used. Instead the mathematicians employ an *intrinsic* standard, *consistency*. Various attempts have been made, all unsuccessful, to extend this standard to the real world. The only result which these attempts have accomplished is to confuse matters and cause an identification of models and the real world.

ROLE OF THE MODEL

The disadvantages inherent in the use of models can be avoided to a large extent by a judicious balancing of the two processes, model-making and data collection. The relationship between these two aspects of Scientific Method deserves careful consideration; it provides one of the main keys to scientific success, and it also involves several notions which can be carried over into our thinking about everyday problems. The relationship can be represented diagrammatically by Figure 1.

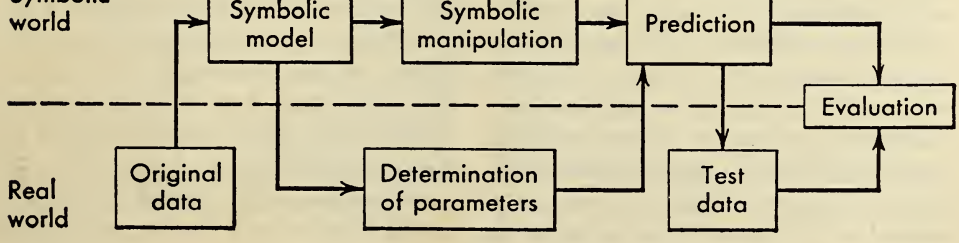


FIGURE 1

The model itself should be regarded as arbitrary; it represents an act of creation like a painting or a symphony. The model can be anything its creator desires it to be. In practice, of course, it is generally stimulated (and therefore affected) by data from the real world (which is labeled "Original data" in Figure 1). Artistic creations also use sensory data. Even in abstract canvases there is some influence from the original data (sensory experience). If the modern artist paints the portrait of a woman, it may not look like a human being to me. But presumably the dabs of paint have some relationship to the woman, though it may require an expert to understand this relationship. Similarly, a physicist's mathematical model of the atom may be far removed from any material substance; again only an expert can appreciate it.

In many cases the symbolic representation used in the model is chosen because it was successfully used in previous models, because it seems plausible to the creator, or because it is convenient. However, some very useful models are based on assumptions which are not evident from common sense or —as in the quantum model—are actually repugnant to common sense.

I would not consider it very plausible to be seated at a desk in Los Angeles and then suddenly to find myself at a desk in Baltimore. It is even less sensible for this jump to have been accomplished in no time at all and without passing through any intermediate point in the process. Yet electrons jump around in this remarkable manner in the quantum theories of physics. Models which embody this curious behavior lead to successful prediction.

Scientists are generally pictured as coldly logical creatures with no disposition to embark on wild flights of fancy. But the geniuses of science have at least as much imagination as any other creative artist. In some respects the symbolic language of science allows greater freedom for expression than the printed word, musical notation, or oil paint.

There is one very important respect in which the scientist differs from the artist, however. The model itself may be arbitrary, but once it is constructed it must meet exacting and carefully specified tests before it is acclaimed as a masterpiece. In the artistic world the criteria for judging the finished product are vague and unsystematic.

There is a second respect in which science and art differ. In art the portrait is the end of the job; in science it is just the beginning. Once the model has been created there are two lines of development—one in the symbolic world and the other in the real world.

In the symbolic world the implications of the model are pursued by manipulations of the symbolic language. If I am interested in the behavior of a pendulum I can set up a mathematical model in which the bob of the pendulum is replaced by a geometrical point. The cord or arm of the pendulum is replaced by a symbol, L, which can be interpreted as the length of the cord. The Newtonian laws may be applied to this model and, by manipulations of the symbolic language, I may derive as a consequence of my model a relatively simple relation between the period (the length of time it takes to complete a

full swing) and the length, *L*. All of this takes place in the symbolic language.

In the real world the numerical value for the length must be obtained. This quantity, *L*, is often called a "parameter." The word "parameter" is merely mathematical jargon for a symbolic quantity, such as *L*, which may be associated with some measurable quantity in the real world. The process of measuring the length of the cord would therefore be called the "determination of the parameter." In most problems there will be more than one parameter involved.

The two paths from the model now join again when the numerical value from the real world is substituted in the formula (derived by symbolic manipulation) in order to obtain the period. The period is found, mathematically, to be proportional to the square root of the length, *L*. If my pendulum is 4 feet long it is easy to calculate that the period will be about 2.2 seconds. This statement is made as a prediction.

In order to test this prediction it is necessary to return once again to the real world. I set up my pendulum and time the swings. I find that the period as determined experimentally is about 2.2 seconds. Perhaps I go ahead and try a whole series of different lengths and the agreement between prediction and experiment seems to be good.

As a consequence of this agreement, I am encouraged to use my mathematical model for prediction purposes and also in the design of clocks or other equipment which utilizes a simple pendulum.

The reader may find it worth while to consider another example, such as the astronomical model of the solar system, and trace through the steps in Figure 1 in order to clarify his own ideas on the role of the model.

One striking characteristic of the relationship between the model and the data is the periodic return to the real world which is indicated in Figure 1. It should be noted that the original data used in the construction of the model may be quite useless for the determination of parameters or testing the model. Hence the return to the real world may not mean merely the collection of additional data, but it may require collection of data of a completely different *type* from the original data.

Now a reader who has forgotten his elementary physics may have wondered why I did not include the weight of the bob as well as the length of the cord in the model of the pendulum. An interesting feature of the mathematical model of the pendulum is that if this additional factor, weight, is included in the symbolic structure, it will cancel out in the manipulations. In other words, the model implies that the period of the pendulum does not depend on the weight of the bob, i.e., the weight is irrelevant in this particular problem. The same thing happens if other factors, such as the way in which the pendulum is set into motion, are included in the model. Thus the symbolic model has served the useful purpose of focusing our attention on the length of the cord. It has therefore suggested an efficient way of experimenting on the pendulum; the *model* has told us what *data* need to be collected.

The little story about the pendulum

had a happy ending, for the model was satisfactory. However, few scientists are so fortunate or clever as to devise a useful model on the first attempt. If prediction from the first model turns out very badly the scientist will have to start over again. The way in which the predictions break down sometimes provides valuable information which can be used to construct a second model.

The role of the model as given by Figure 1 is therefore only a part of a larger sequential process. This sequential role is indicated by Figure 2.

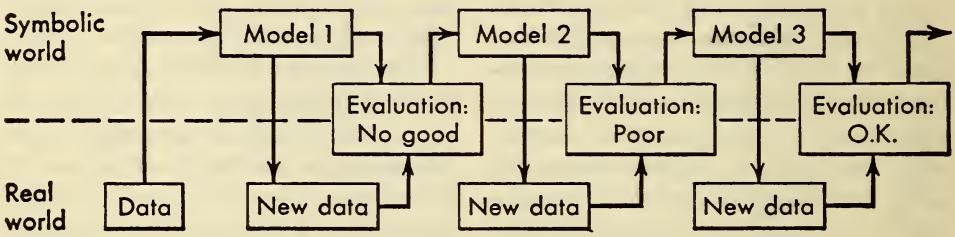


FIGURE 2

The evolution of a successful model generally follows the above pattern. The first shots are often very wide of the mark, but by gradual stages the scientist zeroes in on his target. There is really no end to the sequence. Even after a model has years of successful usage (i.e., Newtonian models in physics), a situation may come along which will not be adequately predicted by the model. A new model must then be developed.

Some readers may find this viewpoint rather unpleasant because they would like this sequence to stop somewhere (i.e., at the truth). Nowhere in the scientific world has this stopping place been attained, although now and then the models have survived for many years. The attitude that the truth had been attained was often a barrier to progress.

A MODEL FOR DATA

The mathematical model for the solar system or for a pendulum can be used for prediction and then tested against actual data. In this test it is not expected that the data and prediction will agree *exactly*. In the pendulum example the predicted period of a 4-foot pendulum is 2.2 seconds. If a 4-foot pendulum is constructed and the period is measured with a stopwatch or other timing device, the periods so measured will be about 2.2 seconds, but there may be some departure from this figure.

Note that these departures of the data from the predicted value have received no allowance in the mathematical model for the pendulum. In order to *evaluate* the model, however, this behavior of the data must be taken into consideration. This may be done intuitively by an argument such as "the departures from the predicted value are very small and quite negligible for practical purposes." A more sophisti-

cated approach is to set up a second model, a model to deal with the measurement data.

Such a model would be a *statistical* model; it would characterize the measurement process itself in mathematical terms. One parameter of this model might be interpreted as the *precision* or repeatability of the method of measurement and this might be estimated from new data collected for this purpose. Many scientific measurements are given in the following form: 2.22 ± 0.10 seconds. The number after the plus-and-minus sign relates to the precision of the measurement. Thus 2.22 might be the average period calculated from a series of measurements on the period of the pendulum. The 0.10 second might indicate that the average is only reliable to 1/10 of a second. We would not be very surprised, therefore, if we had gotten 2.32 or 2.12 seconds as our average period. Consequently, there is no reason to feel that the data contradict our predicted value of 2.2 seconds. If, on the other hand, we had found the average period to be 3.22 ± 0.10 seconds, we would feel that something was wrong either with the model or with the data.

When we set about constructing a mathematical model which will describe data we immediately are confronted with the problem of including, in the mathematical formulation, the well-known inadequacies of data. Thus the inadequacies of the measuring instrument must appear in the model: it must include such things as sensory lapses of the human measuring instrument; various errors introduced by the inanimate instruments as microscopes, telescopes, or clocks; and, in biological work, where an animal is used in the measurement process, all sorts of additional sources of variation due to the animal.

Then there will be incompleteness of the data due to the various steps in abstraction. Some of the data may be irrelevant; some of the relevant factors may have been neglected. Also, only part of the available data may have been collected and only part of this data actually used. In short, any real data will be inadequate and incomplete, and these deficiencies must be included in the model.

It would be hopeless to try to catalogue all the things which might go sour in the process of collecting and utilizing the data, to analyze all of the factors which might operate to influence the experimental results. About all that is possible is to consider broad categories of deficiencies and to include these broad categories in the model.

Now how can these inadequacies, and the resulting uncertainties, be handled mathematically? As you might suspect, this is accomplished by the introduction of the concept of probability into the model. In fact, the notion of probability can be regarded as the distinguishing feature which sets statistical models apart from other mathematical models.

STATISTICAL MODELS

The role of a statistical model is in many respects quite similar to that of any other mathematical model. The diagrammatic representation is indicated in Figure 3. . . .

. . . Occasionally a simple model of this type can be applied to situations in everyday experience. Suppose that I am interested in the proportion of male babies in 10,000 records of live births. There are two outcomes possible when a baby is born (just as in a coin flip)—the baby can be a boy or a girl. I might therefore think of sex determination as analogous to the process of flipping a coin.

One distinction between the coin toss and sex determination is that while the mechanism for determining heads and tails on a coin is fairly well understood, the corresponding mechanism for fixing the sex of a baby is not well understood. Consequently it would be specious to argue that each sex was equally likely. There is, in fact, a large amount of data to show that this is not the case. Hence if a symbol, p, is used in the mathematical model to indicate the probability that a baby will be male, it may not be assumed that $p = 1/2$.

Consequently, one of the things that will have to be done in order to use the model is to obtain data which will enable us to estimate the value of this parameter, p. Perhaps a number such as $p = 0.52$ will be determined from this excursion into the real world.

A second chain of reasoning stays in the symbolic world. Taking the probability as p that a live baby will be a boy, we must answer the question: What will happen in 10,000 births? I will not burden you with the manipulations of probabilities required to answer this question. The mathematics involved in calculating the probabilities for each of the 10,001 possible outcomes becomes too tedious, even for a statistician, and in practice a mathematical approximation which yields useful results with little effort is employed.

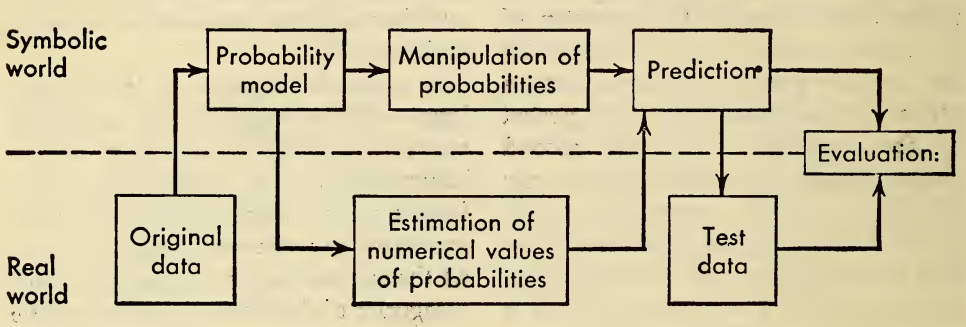


FIGURE 3

With the aid of this device, and substituting the value $p = 0.52$, we can obtain a prediction of the following form: The probability that there will be between 5,100 and 5,300 male births in the sample of 10,000 is equal to about 0.95. In other words, if I am convinced that the model is a good one and that my value of $p = 0.52$ is also reliable, I would be very confident that the actual data should show between 5,100 and 5,300 live male births.

This particular model has taken into consideration only one source of varia-

bility in the data on live birth—the variation due to sampling. Now in practice there are a number of other inadequacies of the data which might very well cause trouble. The reporting procedures may introduce difficulties. In a well-run department of vital statistics in the Western World the tabulation of births may be done rather carefully. On the other hand, if my 10,000 live births were reported by tribal chieftains in a colonial administrative district there might well be a tendency to forget female children.

The problem of *evaluation* of the statistical model is a tricky one. If I found 4,957 boys in the sample of 10,000. I could not say that this result was *impossible* insofar as my model was concerned. The model itself allows a very small chance of this sort of sample.

To a large extent the users of nonstatistical mathematical models can dodge the problem of evaluation by making the evaluation intuitive and simply stating that the agreement of prediction and data is either satisfactory or unsatisfactory. In statistical models one must come to grips with the problem. . . . A major part of a statistician's job lies in the no-man's-land between the symbolic world and the real world, and in particular he must evaluate the predictions of models relative to actual data.

SUMMARY

The key role played by models in scientific thinking is illustrated by several examples. The notion of a model for data is introduced and leads to the concept of a statistical model. The advantages and disadvantages of models are considered. Special stress is laid on the distinction between models of the real world and the real world itself.

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE EVALUATION OF MODELS

KARL W. DEUTSCH

In recent years, increasing attention has been paid to both the use of symbols in the process of thinking, and to the problems that arise when symbols are combined into larger configurations or models—particularly when these models are then used as an aid in investigating or forecasting events that occur in the world outside the thinking system. One important use of such models is in describing the behavior of social organizations.

The organizations to be described may by informal groups, they may be political units or agencies of government, or they may be industrial or business organizations. Each of these organizations is composed of parts which communicate with each other by means of messages; it receives further mes-

"On Communication Models in the Social Sciences," *Public Opinion Quarterly, Fall 1952, 16:3, 356–367.*

sages from the outside world; it stores information derived from messages in certain facilities of memory; and all these functions together may involve a configuration of processes, and perhaps of message flow, that goes clearly beyond any single element within the system. Whenever we are discussing the past or future behavior of such an organization, we must use a model for it, and much of the effectiveness of our discussion may depend upon the degree of similarity or dissimilarity between the model and the thing supposedly modeled.

Investigation of such models, therefore, is more than a mere play upon some fine points in the theory of knowledge. We are using models, willingly or not, whenever we are trying to think systematically about anything at all. The results of our thinking in each case will depend upon what elements we put into our model, what rules and structure we imposed on those elements, and upon what actual use we made of the ensemble of possibilities which this particular model offered.

In one sense the study of models, and the theory of organizations that could be derived from it, cuts across many of the traditional divisions between the natural and social sciences, as well as between the particular social sciences themselves. In all these fields, symbols are used to describe the accumulation and preservation of patterns from the past and their arrangement into more or less self-maintaining, self-destroying, or self-transforming systems. The resulting models are then used to describe further the impact of outside events upon such systems and the responses which each system makes to them. In this manner we use models in describing the behavior of a social group, or of a state, or of a nation, or of the memories and preferences that make up an individual personality. In a similar way, we use models in describing a system of logic, or in suggesting a theory of games, or in describing the behavior of an array of communications machinery.

SOME EARLIER WORK ON MODELS

By a model is meant a structure of symbols and operating rules which is supposed to match a set of relevant points in an existing structure or process. Models of this kind are indispensable for the understanding of more complex processes. The only alternative to their use would be an attempt to "grasp directly" the structure or process to be understood; that is to say, to match it completely point for point. This is manifestly impossible. We use maps or anatomical atlases precisely because we cannot carry complete countries or complete human bodies in our heads.

Each model implies a theory asserting a structural correspondence between the model and certain aspects of the thing supposed to be modeled. It also implies judgments of relevance; it suggests that the particular aspects to which it corresponds are in fact the important aspects of the thing for the purposes of the model makers or users. Furthermore, a model, if it is operational, implies predictions which can be verified by physical tests. A rough survey of major models used in human

thinking in the course of history suggests that there has been a change in the character of the models that predominated in each period, and that it has been a gradual change from pictures to full-fledged models in the modern sense.

THE EVALUATION OF MODELS

We may think of models as serving, more or less imperfectly, four distinct functions: the organizing, the heuristic, the predictive, and the measuring (or mensurative).

By the *organizing* function is meant the ability of a model to order and relate disjointed data, and to show similarities or connections between them which had previously remained unperceived. To make isolated pieces of information fall suddenly into a meaningful pattern is to furnish an esthetic experience; Professor Paul Lazarsfeld once described it as the "Aha!-experience" familiar to psychologists.[1] Such organization may facilitate its storage in memory, and perhaps even more its recall.

If the new model organizes information about unfamiliar processes in terms of images borrowed from familiar events, we call it an explanation. The operational function of an explanation is that of a training or teaching device which facilitates the transfer of learned habits from a familiar to an unfamiliar environment. If it actually does help us to transfer some familiar behavior pattern to a new problem, we may feel that the explanation is "satisfactory," or even that it "satisfies our curiosity," at least for a time. Such an explanation might be subjectively satisfying without being predictive; it would satisfy some persons but not others, depending on each person's memories and habits, and since it yields no predictions that can be tested by physical operations, it would be rejected by some scientists as a "mere explanation" which would be operationally meaningless.[2]

Certainly, such "mere explanations" are models of a very low order. It seems, however, that explanations almost invariably imply some predictions; even if these predictions cannot be verified by techniques practicable at the present time, they may yet serve as *heuristic* devices leading to the discovery of new facts and new methods.[3]

The heuristic function of a model may be independent to a considerable degree from its orderliness or organizing power, as well as from its predictive and mensurative performance.

Little has to be said about the *predictive* function of a model, beyond the well known requirement of verifiability by physical operations. There are different kinds of prediction, however, which form something of a spectrum. At one extreme we find simple yes-or-no predictions; at higher degrees of specificity we get qualitative predictions of similarity or matching, where the result is predicted to be of this kind

[1] Paul Lazarsfeld at a meeting of the Columbia University Seminar on Methods in the Social Sciences, March 12, 1951.

[2] Conant, James B., *On Understanding Science*, New Haven: Yale University Press, 1947; cf. also Bridgman, P. W., *The Logic of Modern Physics*, New York: Macmillan, 1927.

[3] For the concept of heuristics, see Polya, George, *How to Solve It*, Princeton: Princeton University Press, 1944.

or of that kind, or of this particular delicate shade; and at the other extreme we find completely quantitative predictions which may give us elaborate time series which may answer the questions of when and how much.[4]

At this extreme, models become related to measurement. If the model is related to the thing modeled by laws which are not clearly understood, the data it yields may serve as indicants. If it is connected to the thing modeled by processes clearly understood, we may call the data obtained with its help a *measure*—and measures again may range all the way from simple rank orderings, to full-fledged ratio scales.[5]

A dimension of evaluation corresponds to each of these four functions of a model. How great is a model's generality or organizing power? What is its fruitfulness or heuristic value? How important or strategic are the verifiable predictions which it yields? And how accurate are the operations of measurement that can be developed with its aid? If we collect the answers to these four questions under the heading of the "performance" of a model, we may then evaluate the model still further in terms of the three additional considerations of originality, simplicity and realism.

By the *originality* of a model, or of any other intellectual contribution, we mean its improbability. Any idea, scheme or model may be thought of as the product of the recombination of previously existing elements, and perhaps of a subsequent process of abstraction omitting some of the traces of its combinatorial origin. The greater the probability, or obviousness or triteness, of a model, the more frequent is this particular recombination in the ensemble of combinatorial possibilities at the immediately preceding stage. Originality or improbability is the reverse of this value.

A structure of symbols may be highly original but useless. Or a model may be original and perform well but require such a large share of the available means and efforts as to impair the pursuit of other work. Models are therefore evaluated for their *simplicity* or economy of means. But it turns out that the concept of simplicity is not completely simple. Francis Bacon de-

[4] For the relationship of prediction to time series, cf. Wiener, Norbert, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series,* Cambridge: Massachusetts Institute of Technology Press, 1949.

In the natural sciences a yes-or-no prediction might answer a question like this: Will this paper burn or not? A qualitative prediction might answer the question: Will it burn with a bright yellow flame? A quantitative prediction might answer the question: In how many seconds will it heat the contents of a test tube to 400° Fahrenheit?

In economics or politics, yes-or-no questions might be: Will the Jones Corporation build a new plant? Will the Blank party put on a political drive? Qualitative questions might be: Will the Jones Corporation build a large and modern plant? Will the Blank party put on a drive for clean government? Quantitative questions might be: How large a plant will they have built by what date? How many meetings, poster, radio appeals will the Blank party use before next November, and when will the drive reach its climax? It should be remembered that the spectrum formed by these different kinds of questions might well be continuous.

[5] Cf. Stevens, S. S., "Mathematics, Measurement and Psychophysics" in Stevens, ed., *Handbook of Experimental Psychology,* New York: John Wiley, 1951, pp. 1–48.

clared in the controversy between Ptolemaic and Copernican Astronomy that, in the absence of conclusive data from observation, he would choose the simpler of the two hypotheses; he then duly chose the Ptolemaic system on the grounds that it required fewer readjustments of his everyday experience.[6] Clearly, all notions of simplicity involve some sort of minimization problem, but what is to be minimized? Is it the number of unverified assumptions or distinctions, as William of Occam seems to have taught? Or is it a number of calculating steps, as Copernicus suggested in praise of his system? Or is it the number of readjustments of acquired habits, as in Lord Chancellor Bacon's reasoning? If we could succeed in reducing the number of logical or calculating steps required in a model by introducing a large number of suitable fictions, have we simplified the model, or have we increased the complexity of its assumptions? Would we not have simplified it according to Copernicus, but made it more complex according to Occam?

Perhaps the concept of simplicity itself is operational, and could be considered to resemble the concept of efficiency in engineering and in economics. Efficiency in economics denotes the attainment of a given result with the greatest economy in the employment of those means which are shortest in supply at each particular time, place, or situation. Since such supply conditions are historical, simplicity, like efficiency, would then be a historical concept. (If there is merit in this approach, we might wonder about the effect of the availability of cheap calculating aids and electronic calculators on the traditional stress on elegance in mathematics.)

If simplicity is measured by the economy of means in critical supply, then claims to simplicity on behalf of rival models or theories can be evaluated more objectively. We might also be able to predict cross-cultural disagreements about standards of simplicity, as well as changes in accepted standards of simplicity over time. Some of these considerations of simplicity could also be applied to the evaluation of research programs as well as to the measurement of organizational behavior.

The last consideration for evaluating a model or a conceptual scheme is its *realism:* that is, the degree of reliance which we may place on its representing some approximation to physical reality. According to P. W. Bridgman, we may impute "physical reality" to a construct or model if it leads to predictions which are verified by at least two different, mutually independent physical operations. If we put this somewhat more formally, we may say that the statement "X is real" implies the prediction that "Predictions based on the assumption of X will be confirmed by $(2 + N)$ mutually independent physical operations, where N is any number larger than one." The larger N—the number of independent confirmatory operations—actually turns out to be, the greater the degree of reality, or content of reality, we may impute to X. If N approaches infinity, we may be justified in treating X as real, though

[6] Frank, Phillip, *Modern Science and Its Philosophy*, Cambridge: Harvard University Press, 1949, pp. 209–10.

by no means necessarily as exhaustive. This approach implies the assumption that every real object or process is in principle knowable but may be inexhaustible. It may seem farfetched to define the concept of reality as a prediction about a series of other predictions, but it is a definition that can be tested, and I believe, applied to the evaluation of models, or of statements about the inferred inner structure of organizations.

GENUINE VERSUS PSEUDO-MODELS

Mathematical models in the social sciences may lose much of their usefulness through starting from too naive assumptions, or through the introduction of pseudo-constants: that is, magnitudes represented as constants in the mathematical equations, but incapable of being checked by independent and impersonal operations.

An example of sophisticated mathematical techniques prevented from becoming useful by regrettably naive assumptions is found in Professor Nicholas Rashevsky's discussion of changing levels of activity in social groups and of the "interaction of nations," in his *Mathematical Theory of Human Relations.*[7] Professor Rashevsky assumes that members of the politically and economically "active population" differ from the "passive population" by hereditary constitution, and that the relative proportions of "active" and "passive" population then develop according to certain patterns of genetics and natural selection, depending largely on the numbers and density of total population. To what extent Professor Rashevsky's mathematical techniques could be applied to more realistic social and economic assumptions, and particularly to processes of social learning, in contrast to mere heredity, only the future can show.

A far more striking combination of relatively sophisticated mathematics with utter naïveté in social science can be found in the work of the late George Kingsley Zipf.[8] According to Zipf the size of communities in terms of their number of inhabitants should approximate a harmonic series for each country, if its cities were ranked in the decreasing order of size of population. The closeness of the actual distribution found to the theoretical harmonic series was then naively taken as an indicator of social stability. Thus, Zipf found that Austria between the two world wars had too large a capital city and too few cities of middle size, and that the aggregate series of cities in Germany and Austria after Austria's annexation by the Nazis approximated a harmonic series more closely than before. From this he concluded that the German annexations of Austria and the Sudetenland in 1938 had increased the stability of Germany and the social and eco-

[7] Rashevsky, N., *Mathematical Theory of Human Relations: An Approach to a Mathematical Biology of Social Phenomena,* Bloomington, Indiana: Principia Press, 1947, pp. 127–148 and esp. pp. 148–49.

[8] Zipf, George Kingsley, *National Unity and Disunity: The Nation as a Biosocial Organism,* Bloomington, Indiana: Principia Press, 1941; and *Human Behavior and the Principle of Least Effort,* Cambridge: Addison-Wesley, 1949.

nomic balance of her "Lebensraum."[9] This "mathematical" conclusion completely overlooked the fact that before 1938 Germany had already been a food-deficit area, dependent on exports for part of her living, and that Austria as well as the Sudetenland had similarly been areas of food deficits, export dependence, and unemployment. What the Nazi annexations of 1938 had produced had been a merger of three deficits. The "greater Germany" of 1939 was more dependent on food imports and on export drives to pay for them than its component parts; the pooled threats of unemployment in all three territories were met by an armament drive, and food supplies and exports were sought by imperial expansion. What Professor Zipf has described as a harmonic series on paper, was in reality a situation of extreme unbalance and disharmony, which led within a year to a violent explosion in the German invasion of Poland and the unfolding of the Second World War.

Perhaps it is too much to expect at this stage that individuals should undergo the highly specialized training of the advanced professional mathematician and at the same time, the at least equally intense training of the experienced social scientist. The difference in the intellectual techniques in these two fields should not obscure the fact that both approaches represent full-time intellectual jobs. The main task of the mathematician is perhaps to concentrate on the single-minded pursuit of long trains of symbolic operations. He may start out on these from any set of given initial conditions, without caring overmuch, as a rule, why just these conditions or assumptions and no others were selected.

Much of the training of the historian and social scientist is just the opposite. He must become familiar with a very wide range of social and economic situations at different places and times. The outcome of this part of his training is at best a sense of relevance, an experience in judging which factors in a situation must be taken into account and which ones may be neglected without much risk of error. To be sure, the social scientist can only benefit from analytic training. He does and should study economic, political, and psychological theory, and to an increasing extent mathematics and symbolic logic. Yet all analytic work in the social sciences is primarily tied to judgments of relevance, to evaluating the realism of assumptions and the appropriateness of models. This ability is not easily acquired by mathematicians in their periods of rest between or after their more arduous professional labors. And the advice to younger social scientists to study more mathematics should be tempered with the insistence that they will have to judge the relevance of their models against their fund of factual knowledge as social scientists; no amount of mathematical knowledge or advice can take this task from their shoulders.

The most hopeful answer to this problem at the present time lies perhaps in the development of teamwork between men who are primarily social scientists but who have had enough analytical training to put their problems

[9] *National Unity and Disunity*, pp. 196–197 and figure 18.

into a form where mathematicians can go to work on them, and mathematicians who have had enough of a solid training in the social sciences to understand what the social scientists need from them, and how to select lines of mathematical treatment which will lead more closely toward reality rather than away from it.

Another source of trouble with mathematical models in the social sciences stems from the tendency to put arbitrary constants or coefficients into equations so as to make their results fit a known series of numbers or their extrapolations. Thus, Lewis F. Richardson's "Generalized Foreign Politics" attempts to predict the armaments expeditures of two rival countries by equations which contain numerical coefficients for the "grievances" and the "submissiveness" of each country vis-à-vis the other.[10]

It is well known that any finite series of numbers can be fitted by more than one equation, and, on the other hand, that any result can be attained in an equation by introducing a sufficiently large number of arbitrary constants or coefficients. There is all the difference in the world between such arbitrary coefficients and a constant in physics, such as Planck's quantum constant *h*. Genuine constants in physics can be verified by impersonal physical operations of measurement, or by impersonally verifiable inferences from measurement. Such constants are the same for all physicists regardless of their sympathies or political beliefs, and they would be confirmed, in principle, by impersonal recording and measuring devices. The use of such operationally independent and verifiable concepts in models, such as in Bohr's model of the atom, is therefore quite legitimate. As long as social scientists cannot specify an impersonal set of operations for producing a numerical measure of "grievance" or "submissiveness," there will remain a grave suspicion that coefficients based on arbitrary estimates in such matters are somewhat akin to the "variable constants" familiar from the folklore of undergraduate humor.

To be sure, there may be cases where such mathematical pseudo-models may describe, however inadequately, some genuine intuitive insight of their author. It would be folly to suggest that only that is real which is measurable by present-day methods; the perception of *Gestalt* or the structural vision of a previously unrecognized configuration of phenomena all have their places among our sources of knowledge. In all such cases, however, it is the qualitative insights that are relevant, and not the mathematical disguises which they have prematurely donned.

[10] Richardson, Lewis F., "Generalized Foreign Politics; A Study in Group Psychology," *British Journal of Psychology*, Monograph Supplement No. 23, London: Cambridge University Press, 1939; cf. also the summaries in Quincy Wright, *A Study of War*, Vol. II, appendix 42, pp. 1482–83; Kenneth J. Arrow, "Mathematical Models in the Social Sciences," in Daniel Lerner and Harold D. Lasswell, eds., *The Policy Sciences: Recent Developments in Scope and Method*, Palo Alto: Stanford University Press, 1951, p. 137.

✦✦✦✦✦✦✦✦✦✦✦✦ SYSTEMS AND MODELS IN OPERATIONS RESEARCH

OR is primarily a science of method. Its subject matter is the concept of system, rather than the components of that system. Just as a mathematician brings the general formula to bear in a specific case, with the particular values derived from the system under consideration, so the opsearcher brings his systematic background to bear in studying a particular operation. Certainly, the concept of a system is not new in business research. It has been used for many years and by many executives and researchers. But in O.R. some of the implications are taken a bit further and some of the corollaries are made more explicit.

The notion of a system involves the idea of input variables, operating environment, and output results. The easiest of these to consider is the operating environment. This consists of the thousand and one aspects of the actual situation over which, for the purposes of the problem at hand, the executive does not have control. For example, in a problem of optimum inventory policy, the accuracy of the sales forecasting may be considered an environmental variable, while a marketing research study may assume the inventory costs as given. The input variables are those factors of the operation about which the decisions or recommendations are going to be made. They are the variables involved in the hypotheses among which the executive has to choose. A major step in any OR study is made when the input variables have been identified and their range of manipulation determined.

The measurement and analysis of the output variables of a system represent the field where OR has made one of its major contributions to executive action. Often it is quite difficult to state exactly what the purpose of a system is. When you have a rather small system it may seem relatively easy—a man and a lathe are supposed to turn out part A17 for a widget. But what is the compromise between speed and accuracy? How about proper maintenance of the equipment? If he hurries too much he may be careless and have an accident. With a larger system, such as the sales department of a large manufacturing company, the answer to the question "What is the output variable?" may start out as, "To sell more widgets." But this may not be useful or realistic. Many companies measure the sales department's effectiveness in terms of market share, or net profit. And finally

Cost and Profit Outlook, *November 1955, 8:11, 1–4, with permission of Alderson Associates, Inc., Marketing and Management Counsel.*

at the top level within a company, the relevant question is, "Why are we in business?"

The appropriate answers to these questions lead to what opsearchers call a figure of merit. It is the measure of performance that will be used to compare the results of actual or hypothetical changes in the input variables. This figure of merit must satisfy two sometimes divergent sets of criteria. First, it must satisfy the executives who operate the system that it really reflects the overall purpose of the operation. Any situation that they feel is "better" should show a higher figure of merit, and two situations that are about equally good should have approximately equal figures of merit. The second set of criteria are those of usefulness to the Operations Research study. For this purpose, it should be unambiguously defined and should be measurable. If the first requirement is met, the second is often made easier. By borrowing techniques from many fields, and by organizing the experience and insights of the operating personnel, it is usually possible to quantify many factors that seem at first intangible. Morale, worker satisfaction, consumer acceptance, brand loyalty, etc., have all yielded, more or less, to being forced into numerical form. Even when actual numbers cannot be assigned, the variables involved can often be ranked, and modern statistics has developed many powerful techniques for analyzing ranked data.

Having identified the input variables, the operating environment, and the figure of merit, the opsearcher has, in effect, defined his system and has limited what changes can be made in it. This definition can be as narrow or as broad as the research question requires. The virtue of this approach lies not in the size of the problem it can handle (and it has been applied to problems ranging from when to replace light bulbs to how to conduct an entire war) but in the direction it supplies for the gathering and analysis of data, for establishing criteria of relevance for observations, and for constructing and using a model or series of models.

The idea of a model to describe a system is one that is used in many branches of science and has often been used informally in many problem-solving situations. A model of a system is some other system (usually specifically designed for the purpose) which behaves like the system we wish to study. A floor plan of a house or factory is a model since the geometric relations of the drawing parallel the geometry of the actual building depicted. An actual physical model of a bridge or an aeroplane is often used in laboratory investigations of stress and loading. These models give useful results because they behave like the real bridge or aeroplane would. In the actual laboratory analysis, of course, the data from the model have to be interpreted and converted before the results can be applied to the actual situation. This illustrates another aspect of models. Their behavior does not have to appear similar to the system under study. It is necessary, however, that the departures and their laws be known. One of the most useful aspects of models is that while some of their behavior is similar to the system of interest, the rest of the model behavior usually does not

confuse the picture. It is only the relevant behavior that need be similar, and it is the characteristics of the input variables, the operating environment, and the figure of merit that specify the criteria of relevance. There are many sciences that are concerned with human behavior, but each has defined the system differently, so each has different criteria of relevance, and each constructs different models. The medical doctor is looking at a system in which the input variables are age, weight, disease history, etc.; the operating environment consists of diet, amount of exertion, exposure to colds, and other sources of infection, and so forth; and the figure of merit is measured in pulse, respiration, temperature, white corpuscle count, blood sugar level, etc. The lawyer, the market analyst, the economist, the preacher, the psycho-analyst all have very different systems in mind when they look at humans. What is highly relevant behavior for one often is completely inconsequential to the other. To ask which of these systems or models is the right one is to overlook the rule of science that states that the question determines where you look for the answer. Each of the human sciences has different questions to ask so they construct different models. But in each case, the behavior of the model parallels the behavior of the real system in those aspects which are relevant to the questions asked.

The major advantages of a model are obvious. A model can be manipulated more readily than the actual system, even if it is an actual physical model, and much more readily if it is a conceptual model, like a blueprint, or a flow chart, or a mathematical equation. And the manipulation can be done rapidly, easily, and with whatever values for the input variables that are desired. In addition, the results can be observed at once, directly, and without any accidental effects due to the intrusion of the many uncontrollable variables that operate in a real system. Yet, if it is relevant to the problem, these random, unpredictable effects can be put into the model as statistical "noise."

The largest single advantage of a model to the opsearcher and often to the executive in charge of the operation is the insight it yields into the dynamics of the modeled system. Often relationships appear obvious when examining the model that were not even guessed at after prolonged study of the operating system. It is these insights that constitute the real pay-off for Operations Research.

Although models may be of many kinds, the most usual kind takes the form of a mathematical equation. This is not surprising since mathematics is the generalized science of relationships. If the model of a system can be reduced to an equation, many of the powerful tools of mathematical analysis can be brought to bear on it, and its behavior analyzed in minute detail. There has been some criticism directed at Operations Research for its emphasis on mathematics and the charge has been made that it is prone to hide behind complex, involved formulas. This criticism may be justified in some instances, but more often it is the opsearcher's inability to explain that has caused the difficulty rather than the mathematical model itself. For some purposes, a rela-

tively complex model is necessary; for others a very simple equation might do. If the problem involves sales planning in a large company, there may be many variables that should be taken into account, such as salesmen's performance, territory potential, general business trends, technical factors, competitive action and reaction. This may lead to a rather complex model with an impressive looking equation as its statement. An inventory cost model on the other hand, often has only two algebraic terms, one for the portion of the cost that varies with the length of time the item is held in inventory, and one term that represents an average handling cost and is independent of the time. Either may serve the purpose of bringing the pertinent facts to bear on executive decision.

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE DEVELOPMENT AND USE OF MODELS IN OPERATIONS RESEARCH

OR . . . focus[es] the researcher's attention on three major aspects of a situation. These three aspects are first, the input variables, second, the operating environment, and third, the figure of merit. Input variables represent those aspects of the situation under the executive's control. They are the things that the research is going to do something about; they represent the choices of action available to the executive.

The operating environment, however, consists of those factors that do not represent choices of action; at least *for the purpose of the particular investigation* they are to be taken for granted. They define the setting in which the problem exists and include the variables that affect and control the operation under consideration, but whose alterations are not under consideration. However, for the purpose of some other OR study, any of these may become an input variable.

ASKING THE QUESTION

When a problem is first posed, it is frequently expressed in terms of the input variables. How much should we put into television advertising? Should we stock many widgets in a few sizes, or fewer in each size but carry the complete line? How much should we invest in new capital equipment next year? What is the best location for two new warehouses? Executives usually phrase their questions in this form since these questions are all directed toward choices among possible actions. However, the executive will often go fur-

Cost and Profit Outlook, *April 1955, 8:4, 1–4, with permission of Alderson Associates, Inc., Marketing and Management Counsel.*

ther and ask questions not only related to proposed actions, but involving possible outcomes. How will manpower requirements be affected by the proposed promotional campaign? Will the use of a different, more expensive maintenance system reduce the down time of a machine enough to warrant its cost? How accurate should measuring and recording devices be before the value of the added information does not pay for the increased cost of obtaining it?

These questions center attention on the output variables and lead to the third consideration in stating the problem—the figure of merit.[1] This is the quantitative measure of the performance of the system under study. Figures of merit may be simple measures such as total cost per item, rate of production, net profit, number of enemy submarines destroyed, or they may take quite complex forms which involve several factors such as costs, maintenance down time, labor turnover, etc. The figure of merit is the measure it is desired to maximize. When alternative values of the input variables are tested, the values giving the highest figure of merit are considered the best.

DEVELOPMENT OF THE MODEL

Because the definition of the problem, the choice of the figure of merit, and the nature of the model are so interrelated, all three must be developed together. The major problem in model construction, aside from the identification of the variables, is the choice of specific items significant enough to be included. The purpose of the model is to represent the dynamics of the system in a way that is simple enough to understand and manipulate, yet close enough to the operating reality to yield successful results. If the entire system is modeled with every identifiable variable represented in the equation, studying the model or manipulating it may not be much easier than studying the actual system, since the model would then contain extremely complex sets of interrelations and dependencies. It would be most difficult, if not impossible, to secure reliable measures of all of these interrelations. Without such measures, the model would be useless.

LEVELS OF ABSTRACTION

However, it is not the purpose of a model to represent the entire system in its original complexity. The primary value of a model lies in its quality as an abstraction. The more appropriate the abstraction, the more valuable the model. In attempting to find the appropriate level of abstraction the operations researcher is guided largely by the figure of merit. He is constantly asking, "What simplifying assumptions can we make? Will it hurt to drop out differences between salesmen, and use 'average performance'? Can we assume the last five years' trend will continue next year?", etc. To answer these questions often requires experimentation and data collection, or executive judgment may be used to supply the answers. Different models may be developed using different sets of answers. In one model de-

[1] What is here called "the figure of merit" is usually referred to by operations researchers as the *measure of effectiveness.* [Editor]

mand may be assumed to be fixed, known, and continuous. In another model for the same system and problem, the demand may be intermittent and known only as a probability function.

The degree of elaborateness of the model may also depend on the cost of operating with the model as compared with the savings to be made. In a particular inventory situation, a crude, simple model made a 38 per cent improvement in the figure of merit. It was possible to demonstrate that the theoretical maximum improvement was 40 per cent. Obviously it was not worthwhile to pursue the last two per cent.

The choice of the degree of abstraction itself requires an OR approach. If the level of abstraction is too low, i.e., if the model is too much like the system, the advantages of validity are offset by the unwieldiness of the model. If the model is too abstract, the advantages gained from its analytic properties may be more than counteracted by its dubious connection with the actual operating system. In classical economics, the simplifications of perfect competition, complete rationality, full information, and instant action give models which can be manipulated with ease and simplicity, but for many problems their analyses bear little relation to market realities. On the other hand, some of the models often used in market research are so specific and include so many details of the actual situation that they are no more fruitful than the system itself as a source of understanding or control.

In some areas, where much previous work has been done, and where the variables lend themselves to quantification, models at high levels of abstraction are still valid. These areas include inventory control, production lot size, maintenance schedules, etc. In these areas, however, procedures and operations are usually highly efficient, so the contributions of OR, while quite significant and regular, are less spectacular than they are in more dynamic fields of marketing, advertising, sociology, motivation, etc. In these areas operations research has made large but erratic contributions. This is usually a question of the adequacy of the model. If the research team is clever, and extremely lucky, they will develop a model that presents simply and accurately the major dynamics of a complex and poorly understood system. This, of course, yields large benefits. Often, however, they are less fortunate and the model is of a lower level of abstraction, and yields only modest insight and modest returns.

It is constant interplay between the model development and the statement of the problem (particularly the development of the figure of merit), that leads to the best results of operations research in solving the executive problems of business and industry.

✦✦✦✦✦✦✦✦✦✦✦✦✦ SELECTING AN APPROPRIATE MODEL FOR AN OPERATIONS RESEARCH PROBLEM

Any situation in the business world is the result of a bewildering variety of major and minor causes and, unless the less important features are erased from the picture, rigorous analysis becomes hopelessly complicated. The model is the market analyst's substitute for a controlled experiment. For most marketing problems actual experiment is for obvious reasons impractical. The operations researcher therefore lets his mathematics conduct a hypothetical experiment for him. He invents a simplified version of the client's company in which, for example, prices and the business man's advertising expenditure play their normal roles, but minor and unpredictable influences affecting sales are ignored. The mathematics then provides answers to questions about this model. If the model is well chosen these answers will also apply to the facts of the situation.

THE IMPORTANCE OF MODEL CHOICE

The utility of a controlled experiment is largely dependent on the choice of items to be included; that is, the choice of those influences which will be varied so that their effects can be observed and those which will not be permitted to enter. If the trajectory of a missile is much more affected by wind velocity than by the dust content of the air, a controlled experiment which examined only the effects of the former would clearly be far more valuable to a ballistics expert than an experiment which took account of dust to the exclusion of wind. Similarly, the relevance of the operations researcher's models is largely determined by the choice of items to be taken into account and by the manner in which they are included. Unfortunately there is no magic in mathematics. The results of the mathematical computations cannot be expected to be any better than the model on which they are based. It is true that mathematics handled correctly will describe the workings of the model but if the model is poor, results can be irrelevant or even dangerously misleading. The aura of rigor which surrounds an OR analysis must not mislead the business man. He is well advised to observe two important precautions: he should have some degree of assurance that the operations researchers working for him have the experience, skill, and judgment necessary to set up a reasonable model; and as a check, he should examine the logic of the conclusions and recommendations carefully.

Cost and Profit Outlook, *November 1955, 8:11, 1–4, with permission of Alderson Associates, Inc., Marketing and Management Counsel.*

PROCEDURES IN MODEL SELECTION

Since so much is dependent on the choice of a model the procedure to be employed in this selection merits further discussion. Unfortunately, there can be no cut and dried rules for model selection the observance of which will guarantee results. Deciding which factors can safely be ignored is always a matter of judgment since there are usually elements in a business situation which are neither clearly of negligible importance nor clearly crucial. Further, some items may be more important for one problem than another. To give an obvious illustration, warehousing costs can probably be ignored safely in an analysis of a firm's advertising budget but not in an examination of its inventory policies. Again, there are many important borderline cases where the choice is not so clearly indicated. For these reasons the procedures appropriate for model construction can never be detailed to a point where the operation becomes routine and mechanical. The discussion which follows is therefore no more than a listing of general considerations and precautions which can prove useful in the model building process.

STATEMENT OF THE PROBLEM

It is important to begin with a well defined statement of the problem to be considered. This apparently obvious step is too often ignored in practice. Rather vague and hazy discussion of an area to be examined is unsatisfactory. It can easily lead to misunderstanding between the business man and the operations researcher and, more important for the present discussion, it leaves unprovided a piece of information which is indispensable for the choice of model.

This does not mean that the delineation of the problem must be undertaken by the business man. It is often advisable for him to indicate a general area in which, perhaps, he would like improvement in his operations, leaving it to the operations researcher to specify the problem more precisely in the light of his investigations. More often, perhaps, it will be strategic to make the definition of the problem a joint procedure.

Moreover, precise definition of the problem does not imply that the choice of problem, once made, should never be reconsidered. As the analysis proceeds it may become clear that there exist related but more important or more pressing problems.

The need for a clear cut delineation of the problem at hand arises from at least two sources. First, as has already been shown, the importance of a factor will vary with the problem. Whether trucking costs should be explicitly included in the model will depend on whether the problem is one of efficiency in the firm's transportation arrangements or of the organization of its sales force.

Second, the problem will determine the mathematical procedures which are necessary for the analysis and hence the nature of the items which, for mechanical reasons, it would be inconvenient to include in the model. Many

business problems are of the variety which the technician calls optimality problems. For example, "What price for my product will yield the greatest net return?" For this sort of question the techniques of the differential calculus or of the newer linear and non-linear programming are appropriate. On the other hand some business questions seek information only. "What will happen to my sales if I change the color of my package from green to white?" "What will a fall in national income do to my sales?" "What sort of market situation can I anticipate in 1960?" Here different mathematical or statistical techniques will usually be more effective.

LISTING OF RELEVANT INFORMATION

Having specified the problem it is then advisable to turn to an examination of the relevant facts. It is useful to give as complete a description of the situation as is feasible. Again this suggestion is more than a platitudinous injunction to arrive at the truth. Rather it is designed to reduce an unfortunate temptation which besets the mathematical model builder.

The difficulties of statistical fact finding and mathematical computation naturally increase when a more complex situation is investigated. Unfortunately the increase in these difficulties is characteristically far out of proportion to the increase in the number of elements of the situation which are taken into account. As a result the operations researcher is constantly tempted to oversimplify his model. It is all too easy to assume that a computationally troublesome element in the situation is not really very important and consequently can be left out of the model altogether. Neither business men nor operations researchers can always avoid the Freudian pattern of disposing of problems by ignoring them.

For this reason it is very important to have as complete a listing of the facts as possible. Then the decision to leave any element out of the model must be explicit and the model builder is reminded that he must justify these decisions case by case. The danger of careless oversimplification is thereby minimized.

MATHEMATICAL RESTATEMENT OF THE FACTS

Once the facts and the problem have been stated the elements to be included in the model must be chosen. Here there are no general rules at all. Though statistical and other checks on these decisions are sometimes possible, the choices here must ultimately be based on the experience of the firm and the investigator. Every such decision is a matter of balancing the added realism of the model which might be gained by including another element against the cost in increased computational difficulty. Computational complication not only increases the time and money cost of an investigation, but also increases the number of questions which are too difficult to be answered at all with the mathematical techniques currently available.

All these choices having been made the model will be complete when it is

translated into mathematical terms. Though in appearance this and the mathematical computation are the most impressive part of the procedure they are often fairly routine. Where the problem is of a variety which has often been subjected to operations research investigation (as is the case with many pricing problems, product mix necessary to meet specifications problems, inventory problems, etc.) the steps to be followed at this stage are quite familiar and have previously been checked.

THE BUSINESS MAN'S ROLE IN OPERATIONS RESEARCH

Because he often feels himself technically unprepared to examine in detail the mathematical procedures of the operations researcher, the business man often confines his role in an operations research analysis to that of provider of information and recipient of results and recommendations. Of course he should not go too far in the other direction. His own time is too valuable and excessively close supervision in the course of the investigation may make it impossible for the operations researcher to perform his task effectively. But as is the case in every technique operations research can be mishandled. The business man can reduce this danger by careful examination of the final analysis which is usually presented to him. This article has attempted to indicate some of the things to watch for in evaluating their results.

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE ADVANTAGES OF MATHEMATICAL MODELS

ANDREW VAZSONYI

. . . It might be useful at this stage to say a few words about the advantages of using mathematical models. The following list should help:

(a) The mathematical model makes it possible to describe and comprehend the facts of the situation better than any verbal description can hope to do.

(b) The mathematical model uncovers relations between the various aspects of the problem which are not apparent in the verbal description.

(c) The mathematical model indicates what data should be collected to deal with the problem quantitatively.

(d) The mathematical model establishes measures of effectiveness.

(e) The mathematical model explains situations that have been left unexplained in the past by giving cause and effect relationships.

Reprinted with permission from Andrew Vazsonyi, Scientific Programming in Business and Industry, *1958, 18, John Wiley and Sons, Inc.*

(f) The mathematical model makes it possible to deal with the problem in its entirety and allows a consideration of all the major variables of the problem simultaneously.

(g) A mathematical model is capable of being enlarged step by step to a more comprehensive model to include factors that are neglected in verbal descriptions.

(h) The mathematical model makes it possible to use mathematical techniques that otherwise appear to have no applicability to the problem.

(i) A mathematical model frequently leads to a solution that can be adequately described and justified on the basis of verbal descriptions.

(j) It is often the case that the factors entering into the problem are so many that only elaborate data processing procedures can yield significant answers. In such a case, a mathematical model forms an immediate bridge to the use of large-scale electronic data processors.

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE USES AND LIMITATIONS OF MATHEMATICAL MODELS

ROBERT S. WEINBERG

The purpose of this appendix is to discuss the uses and limitations of mathematical models as tools for market planning and to acquaint the reader with the model building process. In the preparation and derivation of market plans, the market planner is faced with many problems associated with the analysis of great masses of interrelated data. These data will consist of facts, estimates, and even guesses. The market planner will generally begin with some information regarding recent past experience, then given an estimate of a future program (which he will relate to this past experience), he will attempt to derive a projection of future requirements or an estimate of the most efficient strategy for attaining a desired marketing objective.

Using mathematical models we are able to construct a framework for organizing this array of data in a manner that will assure internally consistent final results. Our purpose in building mathematical models is to describe the way in which the marketing mechanism operates. If we know the quantitative characteristics of this mechanism (i.e., the interrelation between the various factors and forces which determine company sales or create attractive or unattractive marketing opportunities or situations), we shall be able to project, with a specified level of proba-

An Analytical Approach to Advertising Expenditure Strategy, *1960, 89–116, with permission of the Association of National Advertisers, Inc.*

bility, the inputs, in terms of dollars, physical quantities, man-hours, etc., required to maintain a stipulated level of marketing activity or to attain a given objective.

The relationship between mathematical model building and high speed computing techniques is clear. Before we may consider the application of computers to marketing planning problems, we must first define the problem in some detail indicating a priori, as best we can, all the factors which will influence (or determine) our marketing position; it is not sufficient to merely list the pertinent factors. We must also define the interrelation between these factors. A convenient method, and useful tool, in this connection is the mathematical model. A mathematical model is merely a set of equations, each equation describing the interrelation between the factors (i.e., variables) which determine the marketing mechanism.

In constructing our model we represent the marketing mechanism as a set of interrelated equations explicitly expressing the interactions between the variables which determine the outcome of the marketing situation under analysis. Each marketing or operating relation is represented by one or more equations. For purposes of discussion these equations may be classified according to the nature of the phenomena they describe.

The following four-category classification system is useful:

1) Definitional Equations
2) Technological Equations
3) Behavioral Equations
4) Institutional Equations

The examples outlined below illustrate some of the different problems encountered in developing each type of equation:

1. *Definitional Equations* describe *an exact definitional interrelation* between two or more variables. For example, profits (P) equal sales (S) minus total costs (C), that is $P = S - C$, or residual industry sales (S_I^*) equals total industry sales (S_1) minus company sales (S_c), that is $S_I^* = S_I - S_c$, or dollar sales (S) equals unit sales (U) times average selling price (p), that is $S = U \cdot p$.

2. *Technological Equations* describe the results or interactions of an essentially technological (or physical) process. The production function (from economic theory) relating a company's output (O) to its labor (L) and capital (K) input is a typical technological equation. The equation $O = f(L,K)$ (read as "output equals a function of labor and capital) describes the company's output as a function of labor and capital productivity. For example assume that for a given plant (i.e., a fixed value for K) twenty-five man-hours are required to produce a single unit of output. Further assume that the average worker works forty hours a week and receives two weeks vacation per year.

The annual output of the plant may be represented by the equation:

$$O = 80\ L$$

where:

O = annual output (in units)
L = average number of workers

Note: The value of the coefficient 80 (the average annual output per worker) was computed as follows:

a) Average annual vacation (weeks) $= 2$
b) The number of production weeks per worker per year $= 52 - 2 = 50$
c) Average weekly hours per worker $= 40$
d) Average annual hours per worker $= 40 \cdot 50 = 2000$
e) Manhour labor input per unit of output $= 25$
f) Annual output per worker $= 2000/25 = 80$

The average annual output per worker coefficient shown in the simplified illustrative example developed above reflects the "state of production technology" in the given plant and the resulting productivity of an average worker as measured in terms of output per manhour labor input (i.e., in the present example 0.04 units per manhour labor input, that is, $1/25 = 0.04$).

3. *Behavioral Equations* describe human behavior. For example, the interrelation between consumption (E) and income (Y) or the interrelation between investment (I) and company profits or expected profits (P or $\overline{P}$) could be represented by behavioral equations. Consider two specific examples. The demand for a given consumer good (E_x) may be described by the relation:

$$E_x = 29.3 + 0.000878\ Y$$

where:

E_x = the consumers' total expenditure for the consumer good X (measured in millions of dollars)
Y = disposable personal income (measured in millions of dollars).

In a similar fashion, industry's demand for a given class (product) of capital goods (I_x), and corporate profits (P) may be described by the more complicated equation:

$$I_{x_t} = 6.8 + 0.001216\ P_{t-1} + 0.000624 P_T + 0.002404\ P_{t+1}$$

where:

I_{x_t} = industry's expenditure for producers' durable good X during the *current*, t th years (measured in millions of dollars)
P_{t-1} = corporate profits after taxes during the *previous* $t - 1$, year (measured in millions of dollars).
P_t = corporate profits after taxes during the current, t th, year (measured in millions of dollars)
P_{t+1} = *expected* corporate profits after taxes during the *next*, $t + 1$, year (measured in millions of dollars).

The two illustrative equations shown above represent attempts to statistically describe (i.e., by applying correlation techniques to historical data) two specific human behavior situations, the expenditure patterns of consumers and the investment (in new equipment) decisions of businessmen. For example, if disposable personal income during a given year is forecast to be 381.0 billion dollars, we could expect (on the basis of our statistical analysis of past experience) that the consumer will pur-

chase 363.8 million dollars worth of consumer good X (E_x). [i.e., 29.2 + 0.000878 (381 000) = 29.3000 + 334.5180 = 363.8180]

Similarly if corporate profits after taxes were 19.3 billion dollars *last* year and are forecast to be 27.0 billion dollars *this* year and 31.0 billion dollars *next* year, we could expect industry's current purchases of producer durable good X to be 121.6408 million dollars [i.e., 6.8 + 0.001216 (19 300) + 0.000624 (27 000) + 0.002404 (31 000) = 6.8000 + 23.4688 + 16.8480 + 74.5240 = 121.6408]

It will be noted that in the first equation consumer purchases were directly related to their *current* disposable income. Basic consumption requirements (the level of expenditure that exists independent of the consumer's income) account for 29.3 million dollars in demand and for each million dollars of disposable income consumers will spend an additional 878 dollars for the product. A one billion dollar increase in disposable personal income will lead to an 878 thousand dollar increase in demand.

In the second equation it will be noted that the demand for the class of producer durable goods in question is a function of not one but three variables, last year's profits, this year's profits, and the expected level of next year's profits. This relation assumes that management's decision to purchase new capital equipment will depend (or at least has depended in the past) on three factors: (1) last year's profits insofar as these profits determine the amount of funds management *currently* has available to spend on new equipment, (2) the profits industry expects to earn this year insofar as some of these funds will be available for expenditure (for this or other outlays of higher or competitive priority) this year, and (3) next year's projected profits insofar as these projected profits represent a quantitative index of industry's future profit expectations. The three coefficients (or profit multipliers), 0.001216, 0.000624 and 0.002404, were empirically derived from historic data through the use of multiple correlation techniques. These coefficients are weighting factors based on an objective statistical analysis of past investment behavior. If these weighting factors are accepted, the equation describes an interesting aspect of management's behavior regarding its expenditure for this class of capital goods, that is, last year's profits are almost twice as important as current profits in influencing the investment decision, and expected future profits are almost four times as important.

Behavioral equations describe some historical behavior pattern. Three important points should be emphasized. (1) Unlike the definitional and technological equations developed above (which were exact definitions or descriptions of fairly well-defined or stable essentially physical processes), the behavioral equation is an attempt to develop a mathematical expression describing human behavior. (2) Human behavior often does not fall into precise stable patterns which may be measured statistically; therefore, any forecasts or projections derived from behavioral equations may often be subject to some

(often wide) random error or "shock." Both consumers and businessmen cannot be expected to respond to given stimuli in exactly the same manner in which they have in the past. (3) In using behavioral equations we must recognize that these equations represent a statistical synthesis (or average) of some historic behavioral pattern which may be subject to future change or random shock. In using behavioral equations it is often useful to render the error or shock term explicit, that is, the two illustrative equations discussed above would become:

$E_x = 29.3 + 0.000878\ Y + U_1$, and
$I_x = 6.8 + 0.001216\ P_{t-1} + 0.000624$
$+ 0.000624 P_t + 0.002404\ P_{t+1} + U_2$

The terms U_1 and U_2 are random error or shock factors. . . .

4. *Institutional Equations* describe operating or marketing constraints imposed by internal or external institutional (or policy) factors. Institutional equations can be classified into three categories according to the institutional factors they describe, i.e., internal factors, external factors, joint internal and external factors. Consider the following illustrative examples.

(1) It is the company's policy not to employ more than 5.5 percent of the local labor force in any of its factory areas. This policy imposes a constraint on the maximum size of any of its factories in a given labor market area. That is, in terms of total employment the maximum size of a given factory or company installation may be derived from the institutional equation:

$$W = 0.055\ LF$$

where:

W = maximum total employment in the factory or installation

LF = the total labor force in the specific local labor market area

For example, if the total labor force in a given labor market area is 21,600 workers, the total number of workers the company can employ in this area is 1188 (i.e., (0.055) (21 600) = 1188). If the company employs more than 1188 workers in this area it is violating its own policy.

Other internal institutional equations may be developed describing such policy-imposed operating constraints as product or market (geographic) scopes, fixed minimal debt or other financial ratios, fixed or minimal salesmen to customers (or sales) ratios etc.

(2) External factors such as government taxation policies, government trade regulations and bank lending policies and industry and trade practices will introduce various constraints on the company's marketing operations. These constraints may also be represented by institutional equations. For example, assume that the company's tax rate is 52 percent. Company profits after taxes (P_c^*) can be represented by the institutional equation: $P_c^* = 0.48\ P_c$, where P_c = company profits before taxes. If the company can borrow money (at some specific attractive terms) equal to seventy-five percent of its current assets, (A) the maximum amount of money the company can borrow at these attractive terms (B) may be described by the external institutional equation $B = 0.75$A. That is, if the com-

TABLE 1

THE USES AND LIMITATIONS OF THE FOUR BASIC PHENOMENA EQUATIONS

Equation	*Phenomena Analyzed*	*Inputs Required to Derive Equation*	*Accuracy or Stability of the Equation as an Estimator*	*Major Use of Equation*
Definitional	Exact stipulated interrelations (i.e. exact by definition)	None, represent exact definitions Knowledge of the interrelation between the variables	Accurate by definition—exact by definition	To develop precise interrelations between variables.
Technological	The results or interactions of an essentially technological or physical process	Historical data describing the inputs and outputs of the technological process Knowledge of the dynamics of the process	Normally accurate and highly stable as long as the process has been accurately described and is not subject to sharp change	To forecast or project the results of a technological process in terms of the inputs required to yield a desired output or, conversely, the output that can be expected from a given combination of inputs.

pany has current assets of 644 thousand dollars it can borrow up to 483 thousand dollars (i.e. $0.75 \times 643 = 483$). This equation describes the lending policies of the company's bankers.

(3) The company may have a policy of not borrowing short-term funds if the interest rate exceeds a given amount. However, the company would like to borrow all the money it can up to 75 percent of the value of its current assets at these or better terms. The company's bankers also have a policy limiting the loans they are willing to make at the interest rate the company desires. They ration these loans among their "better customers." To attain a "better customer" status the company must keep a specified minimum cash balance on deposit with the bank. The relation describing the amount of money the customer must keep on deposit in order to maintain its desired borrowing policy in the light of the banks' lending policies would be an example of the third type of institutional equation, describing the impact of both internal and external operating constraints on the company's actions.

Table 1 summarizes the uses and limi-

TABLE 1 (*continued*)

THE USES AND LIMITATIONS OF THE FOUR BASIC PHENOMENA EQUATIONS

Equation	*Phenomena Analyzed*	*Inputs Required to Derive Equation*	*Accuracy or Stability of the Equation as an Estimator*	*Major Use of Equation*
Behavioral	Human behavior patterns; the response of the company's customers, workers, and competitors etc. to a given stimuli	Historical data describing the behavior of a given sector of the population to a given measurable stimuli Knowledge of the behavior pattern	Depends on the persistence and stability of the behavior pattern Subject to random "shocks" since human behavior patterns cannot be perfectly reduced to mathematical equations	To forecast or project the results of a change in some given independent variable (or set of independent variables) on the behavior (i.e. purchasing, expenditure or investment, etc.) pattern of a given human population.
Institutional	Operating or marketing constraints introduced by either the company's own policies or the policies of government, the banking system, or the industry, etc.	Historical or projected values of the institutional constraint parameters Knowledge of the institutional structure	Accurate by definition as long as policy or institutional factors do not change. Subject to error if the projected policies change.	To develop an objective explicit estimate of the impact of a given company or external policy on the company's operations and to render explicit all of the purely institutional constraints which limit the courses of action open to the company.

tations of each of the basic equations outlined above. It should also be noted that in addition to these basic equation types our model may often include several *mixed equations*. A mixed equation may be derived by combining any combination of the four basic equations into a single equation. That is, we may derive a single equation describing both behavioral and technological phenomena. For example an equation describing the rate of market acceptance for a new product could be considered as a joint description of technological change (i.e., new product development) and the adoptive behavior of

the consumer in accepting a new product.

Using the four basic equations as building blocks it is possible to develop up to eleven different mixed equations. The phenomena measured by these equations are as follows (D = definitional, T = technological, B = behavioral, I = institutional):

(1) *DT*	(5) *TI*	(9) *TBI*
(2) *DB*	(6) *BI*	(10) *DBI*
(3) *DI*	(7) *DTB*	(11) *DTBI*
(4) *TB*	(8) *DTI*	

The eleven mixed equation categories listed above merely represent the number of different combinations of phenomena which may be generated from the four basic phenomena. For example, equation (1) describes both definitional and technological factors and equations, (7) and (9) describe definitional, technological, and behavioral factors and technological, behavioral, and institutional factors respectively. Equation (11) describes all four factors. A given mathematical model may combine any number or combination of the fifteen equation types (i.e., the four basic and eleven mixed equations) discussed above.

One important point is worth emphasizing. From a planning point of view **invertibility** is one of the most useful characteristics of mathematical models. Mathematical models are strange sausage meat machines: we can put the sausage in one end, run the machine backwards, and take the live pig out the other end. A mathematical model may (assuming we have preserved its invertibility) be "run" forward or backward. Consider the following oversimplified model.

$$M_t = S_t/100\,000 \qquad \text{(Equation A-1)}$$

$$a_t = 0.06\ M_{t-1} \qquad \text{(Equation A-2)}$$

$$R_t = (M_t - M_{t-1}) + a_t \qquad \text{(Equation A-3)}$$

where:

S_t = company sales during the current (t^{th}) year

M_t = sales manpower requirements during the current (t^{th}) year

M_{t-1} = sales manpower with the company at the end of the previous (t-1) year

a_t = sales manpower attrition during the current (t^{th}) year

R_t = company sales manpower recruitment requirements during the current (t^{th}) year

The three equation sales manpower planning model outlined above may be used to illustrate the important point of invertibility. Equation A-1 is a "sales manpower productivity function." Since on the average one unit of sales manpower is required to produce 100 thousand dollars in sales per annum, the size of the required sales force may be projected by merely dividing expected sales (S_t) by 100 000. Equation A-2 expresses expected sales manpower attrition as a function of the size of the sales force at the beginning of the year (M_{t-1}). In the present case a 6 percent attrition rate is assumed. Equation A-3 is merely a definition of the company's sales manpower recruiting requirements. That is, as long as the company's

sales curve is rising (as long as $M_t > M_{t-1}$), it must recruit enough sales manpower to replace attrition and to support a net increase of $M_t - M_{t-1}$ workers. Now assume that the company starts the year with a sales force of 50 and anticipates sales of 7.5 million dollars during the year. How many sales personnel will the company have to recruit? That is:

GIVEN: $M_{t-1} = 50 \quad S_t = 7{,}500{,}000$
FIND: R_t

$$M_t = 7{,}500{,}000/100{,}000 = 75 \quad \text{(Equation A-1a)}$$

$$a_t = 0.06\ (50) = 3 \quad \text{(Equation A-2a)}$$

$$R_t = (75 - 50) + 3 = 25 + 3 = 28 \quad \text{(Equation A-3a)}$$

ANSWER: the company will have to recruit 28 new sales personnel.

Now we can also run our mathematical model backwards. In the last example we answered the question, given a sales projection how many new people will have to be recruited? Assume that the company believes that it can market all it can produce and the only limitation on sales will be its ability to recruit sales manpower. The company finds that they can recruit 36 new sales personnel. How much can the company expect to sell during the current year? That is:

GIVEN: $R_t = 36$ and $M_{t-1} = 50$
FIND: S_t

$$a_t = 0.06\ (50) = 3 \quad \text{(Equation A-2b)}$$

$$36 = (M_t - 50) + 3, \text{ or} \quad \text{(Equation A-3b)}$$

$$M_t = (36 + 50) - 3 = 83$$

$$83 = S_t/100{,}000, \text{ or} \quad \text{(Equation A-1b)}$$

$$S_t = 8{,}300{,}000$$

ANSWER: the company can expect sales of 8.3 million dollars in the current year.

Given the number of workers the company expects to recruit and the expected attrition rate (from Equation A-2), Equation A-3 may be used to project the sales manpower available during the current year (M_t). M_t may then be substituted into Equation A-1 to derive the final sales forecast.

(Note: The example used above is merely intended to illustrate the invertibility principle and has been oversimplified with two unrealistic assumptions: 1) there is no training lead time required to train newly recruited personnel, and 2) all personnel entering or leaving the company do so on the first day of the year. An actual planning model of this kind would be developed on a monthly or quarterly basis and would have lead and lag relations built in to adjust for the new employee training lead time (i.e., recruitment must precede actual requirement by the required training lead time) and for the month-to-month or quarter-to-quarter changes in personnel joining or leaving the company.)

It is the invertibility principle which makes it possible to compute expenditure planning charts and expenditure planning tables. . . . The invertibility principle is so very obvious and

fundamental that we often neglect to make use of it since we are conditioned to operating the sausage meat machine in only the conventional direction.

Very often relatively simple mathematical models may be used to solve problems that are extremely difficult to solve without the use of such models. Consider the following example involving new product pricing strategy. The company is planning to introduce a new product and desires to set a price that will maximize its profits from this product. As part of the planning process two sets of estimates have been prepared. The company's market research department has estimated the demand for this product at various selling prices and the company's financial department has assembled various estimates of the costs associated with the production and marketing of the product.

Consider these estimates. On the basis of an extensive analysis of the structure of the market the company's market research department prepared the following market forecasts:

Forecast	*Price*	*Estimated Demand*
A	50	7,500
B	100	5,000
C	150	2,500
D	200	0

This market forecast may be used to derive a market demand relation for the new product. This relation describes the interrelation between expected unit sales (u) and the selling price (p). Figure 1 shows the market demand relation derived from the four market forecasts (A through D) above. Figure 1 is a "scatter diagram." Expected unit sales are plotted along the vertical axis and the selling price is plotted along the horizontal axis. For example, the point shown for forecast A is plotted at the intersection of 7,500 along the vertical axis and 50 along the horizontal axis. Similarly forecasts B, C, and D are plotted at the 5,000-100, 2,500-150, and 0-200, intersection points. It will be noted that there is a perfect linear interrelation between demand and price. This interrelation may be represented by the equation:

$$u = 10{,}000 - 50\,p \qquad \text{(Equation A-4)}$$

where:

u = expected unit sales
p = selling price (in dollars)

Consider the derivation of Equation A-4. The demand relation represents demand as function of price. That is, demand is considered as the dependent or endogenous variable and price is considered as the independent or exogenous variable. Demand will vary as price varies. In the present example this interrelation takes a simple linear form. The generalized equation for a straight line is $Y = a + bX$, where Y is the dependent variable and X is the independent variable. That is, Y is a function of X. The exact interrelation between Y and X is determined by the two *parameters* "a" and "b." The param-

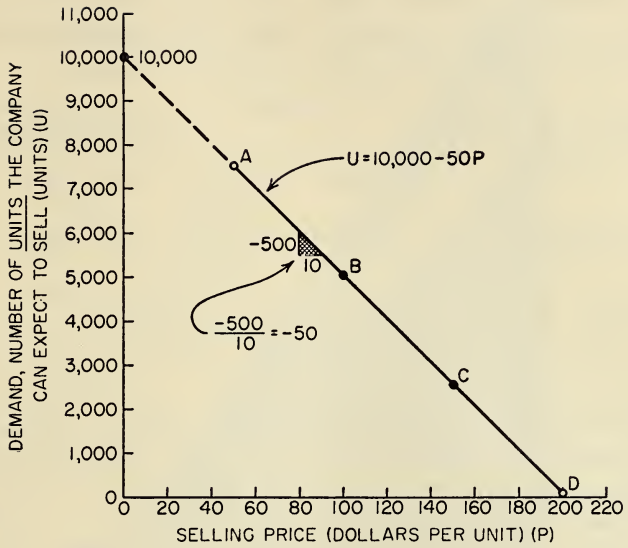


FIGURE 1

THE COMPANY'S NEW PRODUCT DEMAND RELATION

The interrelation between expected unit sales (U) and the new product's selling price (P)

eter "*a*" represents the intercept (i.e., the value of "*Y*" when "*X*" = 0) of the line and the parameter "*b*" represents the slope. Following the extended broken line on Figure 1 it will be noted that at a zero selling price, the company can expect to sell (or rather give away) 10,000 units. While this broken segment of the demand relation only exists mathematically (i.e., as an extrapolation of the solid estimated demand relation) it establishes a value (10,000 units) for the "*a*" parameter of the linear demand relation. To determine a value for the "*b*" parameter or the net change in demand associated with a unit change in price, we merely determine the slope of the line "*b*." This may be determined as follows:

Forecast	*Net Change in Price*	*Net Change in Demand*	*"b"*
A and *B*	+50	−2,500	−50
B and *C*	+50	−2,500	−50
C and *D*	+50	−2,500	−50

In each successive forecast the selling price was *increased* 50 dollars and expected unit demand *decreased* 2,500 units. For each unit (dollar) increase in price, unit demand declined 50 units.

The parameter "*b*" measures the elasticity of demand for the new product. That is, an *increase* of 1 dollar in the product's selling price will be accompanied by a 50 unit *decrease* in demand (i.e., $-2500/50 = -50$). Conversely each one dollar *reduction* in selling price will be accompanied by a 50 unit *increase* in demand. This inverse interrelation between demand and price is represented by the *negative* value for "*b*."

Consider now the new product cost relation. At the request of the financial department the following cost estimates were prepared. The company's manufacturing department estimates that it will cost 50,000 dollars to "tool-up" for production of the new product. They also estimate that *each unit* produced will cost 45 dollars for materials and labor. The marketing department estimates that marketing, transportation, and related costs will be 25 dollars per unit. The financial department estimates that the new product's share of the company's "overhead and burden" will be 35,000 dollars. As shown in Figure 2 the four cost estimates outlined above may be combined to develop the new product's cost-sales relation. The four cost elements fall into two categories, fixed cost elements and variable or unit cost elements:

Fixed Cost Elements		*Variable Cost Elements*	
Tooling	\$50,000	Production (Materials & Labor)	\$45/unit
"Overhead and Burden"	35,000	Marketing, Transportation and Other	25/unit
Total Fixed Costs	\$85,000	Total Variable Costs	\$70/unit

These cost elements may be combined to yield the following total cost (C) — unit sales (u) relation:

$$C = 85{,}000 + 70\ u \qquad \text{(Equation A-5)}$$

Consider Equation A-5 and Figure 2. There is a direct linear interrelation between the total costs (C) and unit sales (u). If total sales are zero the company still incurs 85,000 dollars in total fixed costs and for *each unit* sold, the company spends 70 dollars in production, marketing, transportation, and other variable costs. Figure 2 is self-explanatory. The four cost elements are plotted in cumulative tiers. It will be noted that the two fixed cost elements (tooling and "overhead and burden") are independent of the number of units sold and the two variable cost elements (production costs and marketing and other costs) are directly proportional to the number of units sold. In the present linear cost function the parameter "*a*" = 85,000 dollars and the parameter "*b*" = 70 dollars.

Given Equations A-4 and A-5 the analyst may develop the final pricing strategy model. As indicated above, the

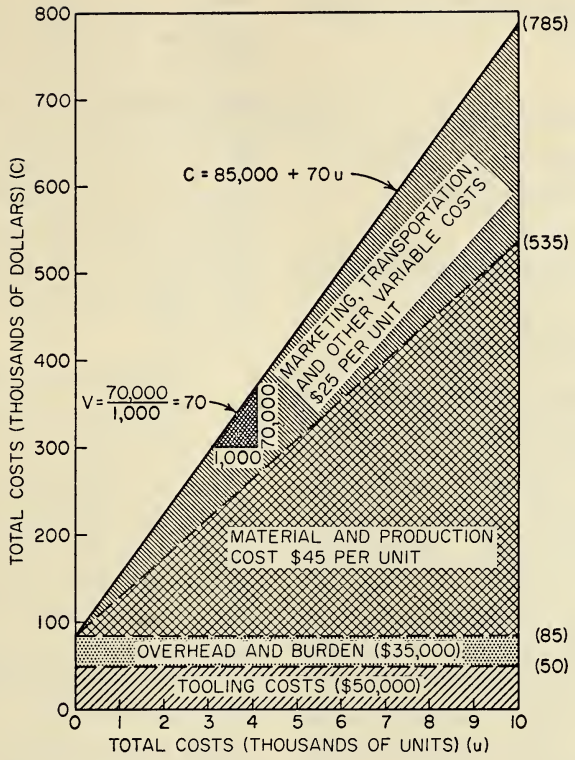


FIGURE 2

THE COMPANY'S NEW PRODUCT COST RELATION

The interrelation between total costs (C) and company unit sales (u)

optimal price for the new product is that price that maximizes the company's total profits. By definition, profits (P) may be derived from the relation.

$P = S - C$ (Equation A-6)

where:

P = Profits (in dollars)
S = Sales (in dollars)
C = Total costs (in dollars)

An estimate of total costs (C) may be obtained from Equation A-5 but an estimate for the company's *dollar* sales (S) is still lacking. Equation A-4 provides an estimate of the company *unit* sales (u) but not its *dollar* sales (S). A fourth, and final, equation is required to complete the model. This equation is merely the definition of dollar sales (S), that is:

$S = u \cdot p$ (Equation A-7)

Equation A-7 merely defines dollar sales as the product of unit sales multiplied by unit price. Combining Equations A-4 and A-7 the following dollar

sales (S) — selling price (p) relation may be derived:

$S = u \cdot p$
(Equation A-7)

$u = 10{,}000 - 50\,p$
(Equation A-4)

$S = (10{,}000 - 50p)p$
(Substituting A-4 into A-7)

$S = 10{,}000\,p - 50\,p^2$
(Equation A-8)

It will be noted that Equation A-8 represents a quadratic rather than a linear relation. This equation is plotted on Figure 3. It will be noted that this curve is a parabola symmetric with respect to the 100 dollar selling price. That is, as the company increases its selling price two things happen: (1) unit demand declines and (2) dollar sales will increase until a maximum point is reached and will begin to decline thereafter. For example, consider the following table:

From the table (computed from Equation A-8) and Figure 3 it will be noted that for any given selling price (p) there is an exact corresponding level for both unit (u) and dollar (S) sales. Now consider any two selling prices p_1 and p_2 and the corresponding unit sales levels u_1 and u_2. Since by definition $S_1 = u_1 \cdot p_1$ and $S_2 = u_2 \cdot p_2$ three basic interrelations follow mathematically:

I: when $p_2/p_1 > u_1/u_2$, $S_2 > S_1$
II: when $p_2/p_1 = u_1/u_2$, $S_2 = S_1$
III: when $p_2/p_1 < u_1/u_2$, $S_2 < S_1$

That is, as long as the ratio p_2/p_1 is greater than the ratio u_1/u_2, S_2 will be greater than S_1. If the ratio p_2/p_2 is equal to the ratio u_1/u_2, S_2 will be equal to S_1. Finally, if the ratio p_2/p_1, is less than the ratio u_1/u_2, S_2 will be less than S_1. In the present example, as long as p_2 is greater than p_1 (i.e., $p_2 > p_1$) and as long as p_2 is equal to or less than 100 dollars (i.e., $p_2 \leqslant 100$) the company's

Selling Price (p)	*Unit* Sales* (u)	*Dollar* Sales* (S)	*Selling Price* (p)	*Unit Sales* (u)	*Dollar Sales* (S)
0	10,000	0	110	4,500	495,000
10	9,500	95,000	120	4,000	480,000
20	9,000	180,000	130	3,500	455,000
30	8,500	255,000	140	3,000	420,000
40	8,000	320,000	150	2,500	375,000
50	7,500	375,000	160	2,000	320,000
60	7,000	420,000	170	1,500	255,000
70	6,500	455,000	180	1,000	180,000
80	6,000	480,000	190	500	95,000
90	5,500	495,000	200	0	0
100	5,000	500,000			

* *Note:* $u = 10{,}000 - 50\,p$
$S = u \cdot p$

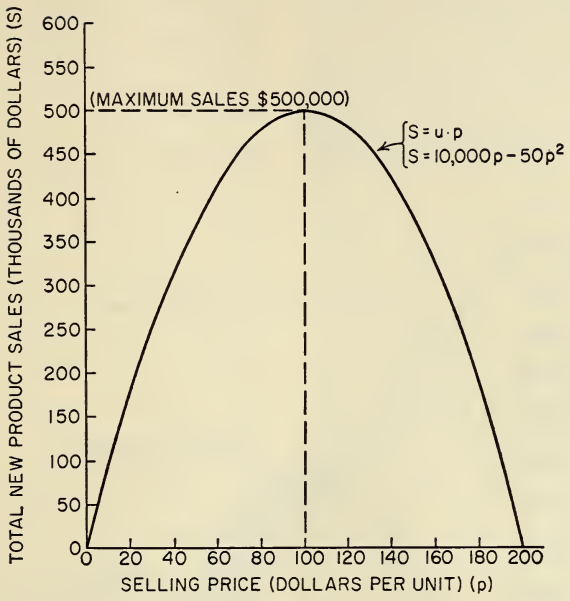


FIGURE 3

THE INTERRELATION BETWEEN NEW PRODUCT TOTAL DOLLAR SALES (S) AND THE NEW PRODUCT SELLING PRICE (p)

total dollar sales will increase as long as p_2 increases. Dollar sales will be at a maximum (500 thousand dollars) when the selling price is 100 dollars. When p_2 is greater than p_1 (i.e., $p_2 > p_1$) and p_1 is equal to or greater than 100 (i.e., $p_1 \geqslant 100$), the company's dollar sales will decline as p_2 increases.

Equation A-6 defines company profits (P) as the difference between dollar sales (S) and total costs (C). Figure 4 shows the interrelation between dollar sales, total costs, and company profits. Two curves are plotted: (1) the solid total dollar sales curve and (2) the broken total cost curve. These curves correspond to the two curves plotted on Figures 2 and 3. It will be noted that two corresponding scales (unit sales and selling price) are shown along the horizontal axis. The common vertical scale and the double horizontal scale make it possible to plot Equations A-5 and A-8 on a single chart. The company's profits are represented by the shaded area between the dollar sales and total cost curves. It will be noted that there are two break-even price levels. A selling price of less than 85 dollars or more than 185 dollars (points A and B) will yield a loss, that is, beyond these points total costs exceed total dollar sales. The optimal profit point is that point at which the dollar sales curve is the greatest distance above the total cost curve. In the present example, maximum profits are 126,250 dollars.

Figure 4 provides a graphic solution to the optimal price-profit maximization problem as long as the distance between the two curves can be accurately measured. The mathematical solution to this problem is actually easier to han-

$$u = 10{,}000 - 50\ p \qquad \text{(Equation A-4)}$$

$$C = 85{,}000 + 70\ u \qquad \text{(Equation A-5)}$$

$$P = S - C \qquad \text{(Equation A-6)}$$

$$S = u \cdot p \qquad \text{(Equation A-7)}$$

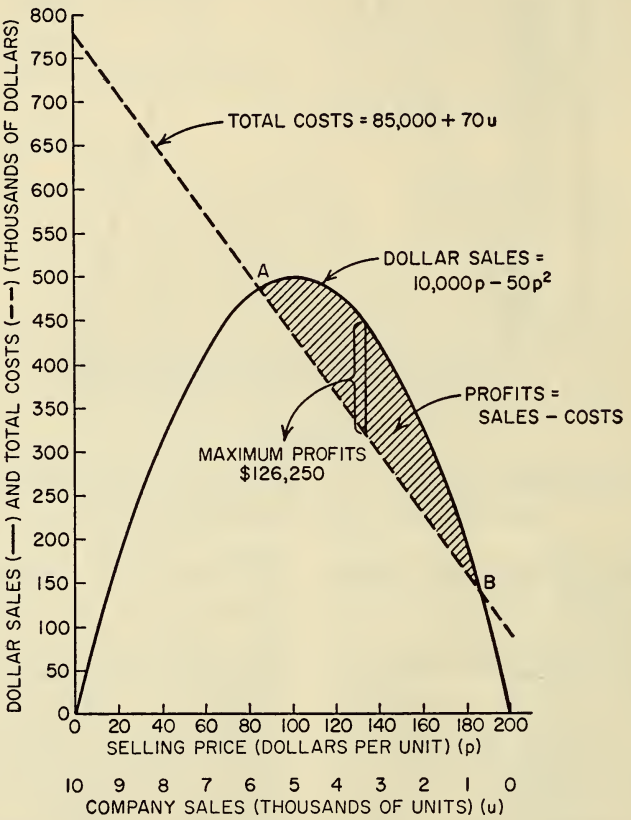


FIGURE 4

THE INTERRELATION BETWEEN DOLLAR SALES, TOTAL COSTS, AND PROFITS

dle than the graphical solution. There are two solutions to this problem: (1) "the marginal cost = marginal revenue solution" and (2) the "profit maximization solution."

The optimal pricing model consists of the following four basic equations:

In review, the present optimal pricing model uses three of the four basic equation types discussed above. The demand relation (Equation A-4) is an example of a behavioral equation describing the expected market reaction associated with changes in the new

product's selling price. The cost-sales relation (Equation A-5) is a technological equation describing the inputs (in dollar terms) required to support a given level of sales (i.e., output). Equations A-6 and A-7 are definitional equations. Equation A-6 defines profits as the difference between dollar sales and total costs. Equation A-7 defines dollar sales as the product of unit sales multiplied by unit selling price. Equations A-6 and A-7 are required to complete the model, i.e., to "close the system." Equation A-7 provides a linkage coupling Equation A-4 which is expressed in *unit terms* with Equation A-5 which is expressed in *dollar terms*. Equation A-6 provides an exact definition of the quantity to be maximized, company profits or the spread between total dollar sales and total costs.

Equations A-4 through A-7 represent the complete pricing strategy model. This model allows the market planner to quickly compute the level of ultimate profits associated with any given new product selling price. The required computational sequence is as follows:

Given "*p*," Determine "P"

(1) *substitute* "*p*" into Equation A-4 to *derive* "*u*"

(2) *substitute* "*u*" into Equation A-5 to *derive* "*C*"

(3) *substitute* "*u*" and "*p*" into Equation A-7 to *derive* "*S*"

(4) *substitute* "*S*" and "*C*" into Equation A-5 to *derive* "*P*"

It will be noted that all four equations are required to complete the computational sequence. If any one of the equations is missing the model is incomplete and would be useless for planning. In the present example the only variable over which the company has control is the selling price (p). That is, the demand and cost relations (Equations A-4 and A-5) are given and assumed fixed. Given the demand situation and its cost structure, the company's only choice of action (once it decides to enter the market) is that of selecting the selling price that maximizes its profits.

As discussed in the text above . . . the company's profits are at a maximum when marginal revenue is equal to marginal cost. Combining Equations A-4 and A-5 the company's total revenue (i.e., sales) may be represented by the equation:

$$S = 10{,}000\,p - 50\,p^2 \quad \text{(Equation A-8)}$$

Similarly combining Equations A-4 and A-5 the following total cost curve may be developed:

$$C = 85{,}000 + 70\,u$$
(Equation A-5)

$$u = 10{,}000 - 50\,p$$
(Equation A-4)

$$C = 85{,}000 + 70\,(10{,}000 - 50\,p)$$
(Substituting A-4 into A-5)

$$C = 85{,}000 + 700{,}000 - 3{,}500\,p$$

$$C = 785{,}000 - 3{,}500\,p$$
(Equation A-9)

The logic of Equation A-9 is clear. If the company's selling price is zero total demand will be 10,000 units (i.e., from Equation A-4, $10{,}000 - 50\,(0) = 10{,}000 - 0 = 10{,}000$). It will cost the company 785,000 dollars to produce and market 10,000 units (i.e., from

Equation A-5, $85{,}000 + 70\ (10{,}000) = 85{,}000 + 700{,}000 = 785{,}000$). From Equation A-4 it will be noted that a one dollar *increase* in price will result in a 50 unit decrease in demand. A 50 unit *decrease* in demand will lead to a 3,500 dollar *decrease* in the company's total costs (i.e., from Equation A-5, $70 \cdot 50 = 3{,}500$). Equation A-9 merely states the simultaneous interrelation between total cost and unit sales and unit sales and selling price in a precise and formal fashion.

. . . the company's marginal or incremental revenue or marginal or incremental cost represents the net addition to the company's *total* revenue or cost resulting from the sale or production of one additional unit of output. Given the company's *total* revenue (sales) or *total* cost relations (equations) the corresponding company *marginal* revenue and *marginal* cost relations (equations) may be readily derived using the basic techniques of differential calculus.*

The company's marginal revenue ($\triangle S$) is equal to the *first derivative* of dollar sales relation; that is

$$S = 10{,}000\ \mathrm{p} - 50\ p^2$$

$$\frac{dS}{dp} = 10{,}000 - 100\ p$$

$$\triangle S = 10{,}000 - 100\ p \quad \text{(Equation A-10)}$$

The company's marginal cost ($\triangle C$) is equal to the *first derivative* of the total cost relation; that is

$$C = 785{,}000 - 3{,}500\ p$$

$$\frac{dC}{dp} = -3{,}500$$

$$\triangle C = -3{,}500 \quad \text{(Equation A-11)}$$

The company's profits will be at a maximum when $\triangle S = \triangle C$, that is, when:

$$\triangle S = \triangle C$$
$$10{,}000 - 100\ p = -3{,}500$$
$$-100\ p = -13{,}500$$
$$p = 135$$

When the new product's selling price is 135 dollars the company's unit sales will be 3,250 units (i.e., $10{,}000 - 50(135) = 10{,}000 - 6{,}750 = 3{,}250$) and its dollar sales will be 438,750 dollars (i.e., $135(3{,}250) = 438{,}750$). When the company's sales are 3,250 units its total costs will be 312,500 dollars (i.e., $85{,}000 + 70(3{,}250) = 85{,}000 + 227{,}500 = 312{,}500$) and its total profits will be 126,250 dollars (i.e., $438{,}750 - 312{,}500 = 126{,}250$). A similar computation demonstrates that this is the optimal selling price-profits combination.

* *Note:* The basic concept in differential calculus is that of the derivative (or more strictly expressed, the derivative of a dependent variable "*y*" with respect to an independent variable "*x*") which is generally represented by the compound symbol *dy/dx*. It is important to remember that this is a compound symbol, that is, a single symbol meaning the derivative of "*y*" with respect to "*x*". It is *not* "*dy*" divided by "*dx*" or "*d*" multiplied by "*y*" divided by "*d*" multiplied by "*x*." In precise mathematical terms a derivative of a function is the limit of the ratio of the increment of the function to the increment of the independent variable when the latter increment varies and approaches zero as a limit. For the purposes of the present discussion, however, this mathematical definition may be replaced by the less rigorous definition, the derivative of "*y*" with respect to "*x*" is the instantaneous rate of change of "*y*" with "*x*." In the present example we are concerned with the rate of change in company revenue (sales) (S) and in company costs (C) and later in company profits (P) associated with a unit change in the company's selling price (p).

Selling Price (dollars)	*Unit Sales (units)*	*Dollar Sales (dollars)*	*Total Costs (dollars)*	*Total Profits (dollars)*
134	3,300	442,200	316,000	126,200
135	3,250	438,750	312,500	126,250 (optimal profits)
136	3,250	435,200	309,000	126,200

The second approach to the optimal pricing-profit maximization problem consists of maximizing the new product total profit relation itself. That is, combining Equations A-4, A-5, A-6 and A-7 the following total profit relation may be developed:

$$P = S - C$$

(Equation A-6)

$$P = (10{,}000\ p - 50\ p^2) - (85{,}000 + 700{,}000 - 3{,}500\ p)$$

$$P = (10{,}000\ p - 50\ p^2) - (785{,}000 - 3{,}500\ p)$$

(from Equations A-8 and A-9)

$$P = 10{,}000\ p - 50\ p^2 - 785{,}000 + 3{,}500\ p$$

$$P = 13{,}500\ p - 50\ p^2 - 785{,}000$$

(Equation A-12)

$$\frac{dP}{dp} = 13{,}500 - 100\ p = 0$$

(Equation A-13)

$$100\ p = 13{,}500$$

$$p = 135 \text{ dollars}$$

As demonstrated above the company's optimal selling price is 135 dollars.

Figure 5 shows a graphic solution to the "marginal revenue = marginal cost" approach to the profit optimization problem. The solid line represents the company's marginal revenue relation (Equation A-10) and the broken line represents the company's marginal cost relation (Equation A-11). These two relations intersect (i.e., marginal revenue = marginal cost) at a point where the selling price (p) equals 135 dollars.

Figure 6 shows a graphic solution to the "profit maximization" approach to the profit optimization problem. The solid curve shows the company's profits-selling price relation (Equation A-12). This relation, like Equation A-8, is a quadratic equation; that is, a parabola symmetric with respect to the optimal selling price. The company's profit reaches its peak (126,250 dollars) when the selling price is 135 dollars.

Consider Equation A-12. In our discussion of the derivation of Equations A-8 and A-9 we demonstrated that the company's dollars sales (S) and total costs (C) may be represented as functions of the new product's selling price (p). Equation A-12 was derived by merely substituting Equations A-8 and A-9 into Equation A-6. It follows that the company will earn a profit as long as $13{,}500\ p$ is greater than $50\ p^2 + 785{,}000$ or as long as $13{,}500\ p$ is at least 785,000 dollars greater than $50\ p^2$. From Figure 6 it will be noted that as the selling price increases the company's profits increase, reach a peak, and decrease thereafter. The company's profits are represented by the shaded area of the curve above the zero or break-even profit axis. The company can increase its profits by increasing its selling price until the optimal selling price (135

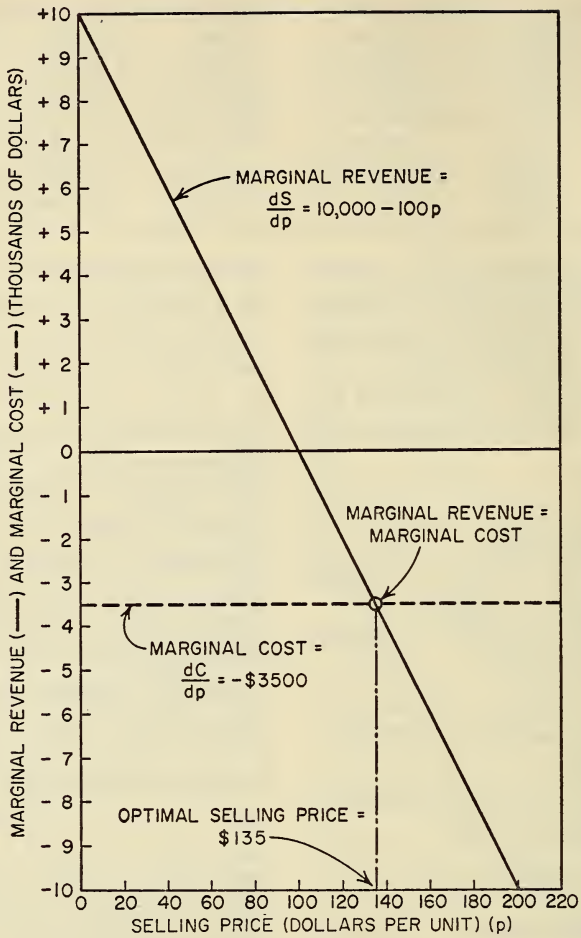


FIGURE 5

A GRAPHIC SOLUTION TO THE OPTIMAL SELLING PRICE—MAXIMUM PROFITS PROBLEM: MARGINAL REVENUE—MARGINAL COST APPROACH

dollars) is reached. As the company sets its selling price above this level its total profits will decline. A selling price in excess of 185 dollars will yield a loss. A selling price of less than 85 dollars will also yield a loss. If the company sets a selling price of *less than 135 dollars* it is *underpricing* and, *under the conditions assumed above* (*i.e., Equations A-4 and A-5*), is foregoing a potential profit opportunity since the resulting increase in rate of profit on those units sold will offset the decline in demand associated with the price increase (note: this situation is critically determined by the shape (slope)

of the company's demand relation). Conversely, if the company sets its selling price *above 135 dollars* it is *overpricing* and is foregoing a potential profit opportunity since the decline in unit demand associated with the higher price *is not* offset by a corresponding increase in the rate of profit on those units sold.

One of the most powerful uses of mathematical models is in the area of deriving generalized basic market planning formulae. Consider a generalized solution to the problem discussed above. The basic relations developed above may be used to derive a generalized pricing formula applicable to any situation in which we have a linear

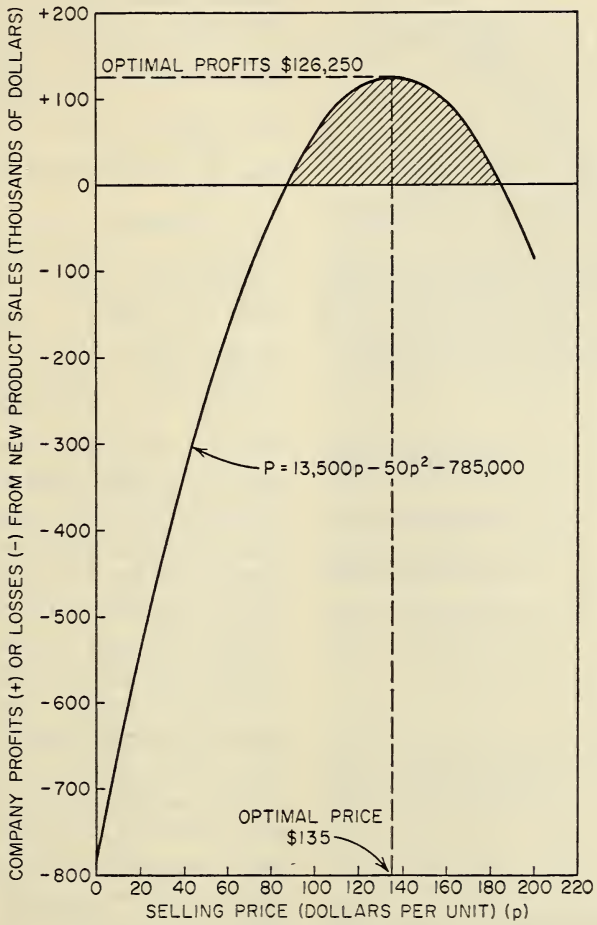


FIGURE 6

THE INTERRELATION BETWEEN COMPANY NEW PRODUCT PROFITS (P) AND THE NEW PRODUCT'S SELLING PRICE (p)

demand relation and a linear cost relation. This formula may be derived as follows:

Let:

P = company profits (in dollars)
C = total costs (in dollars)
S = sales (in dollars)
u = sales (in units)
p = selling price
f = total fixed costs (in dollars)
v = total variable costs (in dollars per unit)
b = the intercept of the demand relation (i.e., unit sales when $p = o$)
e = the slope of the demand relation (i.e., the elasticity of demand)

Then:

$P = S - C$ (Equation A-6a)

$S = u \cdot p$ (Equation A-7a)

$C = f + v \cdot u$ (Equation A-5a)

$u = b + e \cdot p$ (Equation A-4a)

It follows that:

$S = (b + ep)\, p$
(Substituting Equation A-4a into Equation A-7a)

$S = bp + ep^2$ (Equation A-8a)

$C = f + v(b + ep)$
(Substituting Equation A-4a into Equation A-5a)

$C = f + vb + vep$ (Equation A-9a)

$P = (bp + ep^2) - (b + vb + vep)$
(Substituting Equations A-8a and A-9a into Equation A-6a)

$P = bp + ep^2 - b - vb - vep$
(Equation A-12a)

Equation A-12a is a generalized version of Equation A-12 above. This equation is a quadratic of the type shown on Figure 6. As shown above the company's profits will be maximized when we select a value for "p" that sets the first derivative (i.e., $\frac{dP}{dp}$) of this relation equal to zero. It will be recalled that $\frac{dP}{dp}$ measures the change in profits (P) associated with a given change in price (p). The change in profits associated with a given change in price may be positive, negative, or equal to zero. Now if the company chooses a price that is *less than optimal,* an increase in price will yield an increase in total profits and a decrease in price will yield a decrease in profits. Conversely, if the company chooses a price that is *greater* than optimal an increase in price will yield a decrease in total profits and a decrease in price will yield an increase in total profits. *There is only one point at which both an increase or decrease in price will yield a decrease in total profits.* This point is the optimal price point, i.e., 135 dollars in the example developed above. Assume that the company sets an initial price somewhere below the optimal price. As the company increases its price its total profits will increase until the optimal price is reached. If the company continues to increase its selling price above the optimal price its total profits will decline. As the price increases the sign of the net *change* in total profits will change from positive to negative. Total profits will be at a maximum at that precise point where the net change in total profits is nei-

ther positive nor negative, i.e., is zero.

The first derivative of Equation A-12 is:

$$\frac{dP}{dp} = b + 2ep - ve \quad \text{(Equation A-13a)}$$

Setting Equation A-13a equal to zero and solving for "p"

$$\frac{dP}{dp} = b + 2ep - ve = 0$$
$$-2ep = b - ve$$
$$2ep = ve - b$$
$$p_o = \frac{v}{2} - \frac{b}{2e} \quad \text{(Equation A-14)}$$

Where: p_o = the optimal selling price (optimal value for p).

Equation A-14 represents a generalized solution to the pricing strategy problem discussed above. This equation expresses the optimal price for a new product as a function of three variables: (1) the total variable cost rate (v) (2) the intercept of the demand relation (b), and (3) the slope of the demand relation (e). The important point is that Equation A-14 is applicable to *any* pricing situation in which the demand and cost relations are linear. (Note: The same basic techniques may be used to develop similar pricing models for nonlinear cases.) Equation A-14 may be used to solve the specific example discussed above. In this example the following values were assumed:

$v = 70$
$b = 10{,}000$
$e = -50$

Substituting these estimates into Equation A-14 yields the following results:

$$p_o = \frac{v}{2} - \frac{b}{2e}$$
$$p_o = \frac{70}{2} - \frac{10{,}000}{2(-50)} =$$
$$p_o = \frac{70}{2} - \frac{10{,}000}{-100}$$
$$p_o = 35 - (-100)$$
$$p_o = 35 + 100 = 135$$
p_o = optimal price = 135 dollars.

The use of mathematical models as tools for market planning may be regarded by some as "nothing more than some long-hair gadget or impractical academic exercise." This is in part the fault of the "impractical" mathematical model builder, who loses sight of the fact that the model is a means to an end and becomes hypnotized by the muse of mathematics and so involved with mathematical notation, niceties, and technique that the model itself becomes the end and not the means. If a mathematical model is to be kept a useful tool assisting the planner to reach a conclusion, it must be kept as simple as possible, describing the operating structure it represents in a simple and concise manner (that is, in a form as simple and concise as possible). The analysis of complicated problems will always require complicated models.

When the uses or applications of a mathematical-model approach are discussed, four arguments against their use are generally offered:

(1) The Fallacy of "Argumentum Ad Hominen":

This argument attacks the results obtained by some unsuccessful model builders. This is an instance of the clas-

sic fallacy, the invalidity of a method is not necessarily proven by pointing to the error or weaknesses of those individuals who follow (or claim to follow) the method.

(2) The Closed System Argument:

This is the familiar "How can a mechanical or semi-mechanical procedure allow for the important factor of mature human judgment?" Argument: The mathematical model is a tool designed to assist the planner in reaching a conclusion; therefore, it does not provide, nor is it intended to provide, a final judgment-free mechanical answer (see advantage 1 below).

(3) The Inadmissibility of Technique and Technical Terms:

The mathematical model-builder uses terms and techniques (from the fields of mathematics, statistics, and econometrics) with which the average man is unfamiliar and therefore, suspicious of. This, however, is a weakness common to any specialized technical field. Since the model-builder uses terms and a "shorthand" with which the average person is unfamiliar, there is a marked tendency for the layman to consider him as "impractical or too theoretical."

(4) The Past, Present and Future Argument:

This is the popular misconception that the hypothesis of the future cannot be grounded on the past and the present. It is argued that in a dynamic system the interrelationships between factors are changing so rapidly that it is almost impossible to project future trends on the basis of any form of purely mathematical extrapolation of past experience. This, however, is not our intention; the model is not intended to be used as an "automatic mechanical robot estimate generator" but rather as a tool for data analysis. The model helps us systematically organize our data for analysis (see advantages below). The model helps us to more fully define and utilize our past experience in the most efficient manner. How can we ever plan for the future if not on the basis of our past experience?

In using a mathematical model as a tool for market planning, five worthwhile advantages are gained. The first three are general advantages, the last two are technical (computational) advantages:

(1) The model constitutes a well-defined statement of the problem. As such, it enables us to state, and to employ, all of our a priori information on the problem at hand.

(2) The model renders explicit the assumptions on the basis of which the investigation or analysis proceeds. That is to say, we incorporate our basic assumptions right into the model itself.

(3) The model makes it possible, if it is possible at all, to orient our research to answer specific questions of policy.

(4) From a mathematical and statistical standpoint the model, itself, when fully formulated, serves to determine the statistical techniques which ought to be employed so that (a) there are no inherent contradictions in our procedure and (b) our estimates have certain technically desirable properties (they

are consistent, they have known error probabilities, etc.).

(5) Once we arrive at a set of estimates the model enables us to understand and interpret them without difficulty, as long as we have selected our statistical (estimating) technique properly. The final results are readily reproducible and lend themselves to an objective scientific evaluation.

Mathematical model building is a four-stage, and often an experimental process:

(1) The Formulation Step—Study the marketing situation under consideration, listing all the factors which influence the "outcome" and the interrelations between these factors. Develop a flow chart or "qualitative model" of the situation. Convert the "qualitative model" into a generalized mathematical model.

(2) The Data Collection Step—Collect the basic data required to derive empirical estimates of the parameters indicated in the generalized model.

(3) The Computation Step—Compute the parameters required to convert the generalized model developed in Step 1 into a specific planning and control model.

(4) The Verification and Adjustment Step—Compare the estimates generated from the model with their actual counterparts and thus verify the model. If the estimates deviate from their actual counterparts, determine the adjustments and/or changes which must be made in the model in order to eliminate these deviations.

In closing . . . , it should again be emphasized that our purpose in building a mathematical model is to describe the way in which the marketing mechanism operates. The model is nothing more than a tool; it is in many respects analogous to a skeleton on which we may hang our data for systematic analysis. Mathematical models are not intended as a substitute for marketing judgment, but they can be used to narrow the range over which marketing decisions must be made on the basis of judgment alone. While mathematical models will never yield panaceas or a set of "cure-all" formulas, they provide a logical means for utilizing *all* available information to develop more realistic and efficient marketing plans.

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE USE OF MATHEMATICS IN PRODUCTION AND INVENTORY CONTROL

ANDREW VAZSONYI

In the course of studying production and inventory control, the author of this paper found it repeatedly necessary to explain his mathematical theory to operating personnel, many of whom have had little or no

Management Science, *October, 1954, 1:1, 70–85.*

formal training in mathematics at the college level. This paper, then, is an outgrowth of these presentations and has a twofold purpose: (1) to present a theory of a very small part of the problem of production and inventory control with the objective of acquainting the reader with the nature of the mathematical methodology, (2) to lay emphasis on a didactic presentation to show the principles involved in explaining these mathematical concepts.

The subject matter to be discussed is what is generally called the preparation of parts requirement lists and production explosion charts.

In order to fix ideas, we paraphrase the problem in question as follows:

> The manufacturing planning program begins with breaking down the sales forecast into the requirements for the detailed subassemblies and parts. A production explosion chart is plotted showing, for a manufacturing unit, first, the breakdown into major assemblies and separate parts, and then at each successive step the further breakdown into subassemblies and parts. Based on this "explosion system" the requirements of all assemblies and subassemblies are determined before any parts manufacturing orders are initiated.[1]

Every production man knows that a great deal is implied in the above statement; however, the statement contains little clue as to what these implications are. The fact of the matter is that many pages of verbal discussions of the subject are required to develop these implications and in many instances, from the point of view of operating personnel, there is no complete description available in the literature at all. In contrast, we proceed here to develop a statement in terms of mathematical concepts which not only suggest procedures and methods but in fact specifically contain answers to some of the fundamental issues involved in the above statement.

In order to develop such a theory it is necessary to examine these questions in detail and to, step by step, build a bridge from this above statement to mathematical equations. That such a thing can be done and that it can be useful, most likely is a perplexing thought to the reader.

THE CONCEPT OF THE ASSEMBLY PARTS LIST

We begin by considering Figure 1 where a sample assembly parts list is shown—each assembly to be made has a similar sheet. The assembly in question is a "Panel" with Part No. 435090-012. This number is shown on the upper right-hand corner under the word "Makes Assembly." This assembly is made up of seven different articles[2]—bushing, panel black, etc. The part number of each of these is given on the sheet. Under the heading N. A. QTY. (next assembly quantity) it also is shown how many of these articles are needed. Thus, the bushing with Part No. 420990309 is required in a quantity of three for each of the Panels 435090-012.

The Assembly Parts List has some other information which for the purpose

[1] *Production Handbook*, L. P. Alford and J. R. Bangs, p. 230, "Breaking Down Requirements."

[2] An "article" might be an assembly, subassembly, or part.

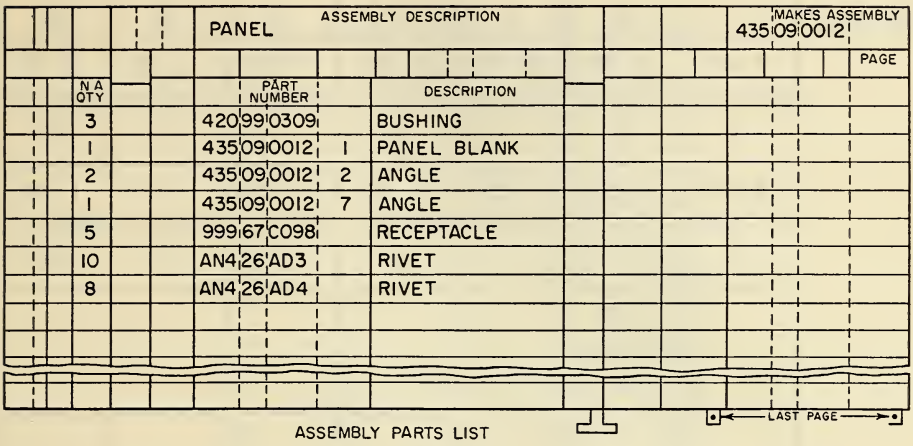


FIGURE 1

This sheet refers to the assembly "panel" which is assigned the part number 435090012. This panel is made up of seven different articles which could be subassemblies or parts. The panel requires three bushings 420990309, and one panel blank 435090012-1, etc. Each assembly will have a similar sheet. The type of information contained on these sheets will form the basis of our discussion.

of our simplified discussion is disregarded.

In order to build a mathematical model, the information in the Assembly Parts List is to be put in a concise symbolic form. Note first that the information in Figure 1 is redundant; the panel in question has the Part Number 435090012 and, therefore, this assembly could be identified solely by this number.

ARTICLE		435090012
3 OF	ARTICLE	420990309
1 OF	ARTICLE	435090012-1
2 OF	ARTICLE	435090012-2
1 OF	ARTICLE	435090012-7
5 OF	ARTICLE	99967C098
10 OF	ARTICLE	AN426AD3
8 OF	ARTICLE	AN426AD4

FIGURE 2

ABBREVIATED ASSEMBLY PARTS LIST

The names of the various articles are omitted.

Figure 2, then, presents an abbreviated parts list; the names of the articles are no longer listed.

In the remainder of the report, for the sake of brevity, we shall not carry these long part numbers, but shall assume that the articles are numbered 1, 2, 3, etc. With these short-cuts, then, a set of assembly parts lists might take the form of Figure 3. This purely hypothetical manufacturing process (which bears no relationship to the one described in Figures 1 and 2) deals with Assemblies 1, 2, 4, 5, 7, 8, and 9. As-

ARTICLE	1	2	4	5	7	8	9
	1 OF ARTICLE 3	2 OF ARTICLE 6	2 OF ARTICLE 1	3 OF ARTICLE 3	1 OF ARTICLE 1	1 OF ARTICLE 1	3 OF ARTICLE 6
	2 OF ARTICLE 5	1 OF ARTICLE 7	1 OF ARTICLE 7	1 OF ARTICLE 6	2 OF ARTICLE 5	1 OF ARTICLE 5	1 OF ARTICLE 8
		2 OF ARTICLE 8					

FIGURE 3

SET OF ABBREVIATED ASSEMBLY PARTS LISTS

Each column represents a single assembly parts list. For instance, article 2 is made up of two of article 6, one of article 7, and two of article 8.

sembly 1 is made up of one of article 3 and two of article 5. Each column of Figure 3 represents a single assembly parts list similar to Figures 1 or 2.

A further saving of words can be effected by saying A instead of article and saying A_1 for article 1, A_2 for article 2, etc. With this notation then Figure 3 becomes Figure 4.

The information in Figure 4 can be presented in a somewhat different form using rectangular tables. Such a table is shown in Figure 5. The information is the same as in Figure 4; for instance, it can be seen that A_4 is made up of two A_1's and one A_7. The advantage of the new presentation is that it is conceptually more descriptive. One can

A_1	A_2	A_4	A_5	A_7	A_8	A_9
$1A_3$	$2A_6$	$2A_1$	$3A_3$	$1A_1$	$1A_1$	$3A_6$
$2A_5$	$1A_7$	$1A_7$	$1A_6$	$2A_5$	$1A_5$	$1A_8$
	$2A_8$					

FIGURE 4

SET OF ABBREVIATED ASSEMBLY PARTS LISTS

The word Article 1 is replaced by A_1, Article 2 by A_2, etc.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
A_1				2			1	1	
A_2									
A_3	1				3				
A_4									
A_5	2						2	1	3
A_6		2			1				
A_7		1		1					
A_8		2							1
A_9									

FIGURE 5

TABLE OF ASSEMBLY PARTS

The information still is the same as on Figure 4 but the presentation is more systematic. Note that for completeness *all* the articles are listed in the top now. For instance, A_3 requires no other article as A_3 is not an *assembly* but a *part*.

talk about a column, describing what an assembly is made of, or one can talk about a row, showing what an assembly goes into. For instance, A_1 is made of one A_3 and two A_5's; A_5 goes into A_1 twice, into A_7 twice, into A_8 once, and into A_9 three times.

The same information is finally presented again in Figure 6. What we did is very simple; we filled in the empty squares with zeros and omitted the A's. This presentation is very simple because we can say that every article goes into every other article—the numbers on the table show how many times. The zeros mean that a particular article does not "really" go into the other article in the ordinary sense; from our point of view, this distinction need not be made.

For explanation, let us insert an analogy from algebra. We can subtract any two numbers, five less five equals zero. If we did not have zeros we would have to say that subtraction can be carried out only under certain circumstances. We always would have to watch that the formulas have meaning. Therefore, the invention of the zero is an extremely useful thing. It also forms the essential foundation of the concept of Arabic numbers as distinct to Roman numerals where zeros do not exist. Anyone who would have the courage to carry through a division in Roman numerals would appreciate the point.

One further point—we have put zeros into the "diagonal" elements [3] of the table. The question of how many A_2's

[3] The diagonal elements are formed by the first number of the first row, the second number of the second row, the third number of the third row, etc.

	1	2	3	4	5	6	7	8	9
1	0	0	0	2	0	0	1	1	0
2	0	0	0	0	0	0	0	0	0
3	1	0	0	0	3	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	2	0	0	0	0	0	2	1	0
6	0	2	0	0	1	0	0	0	3
7	0	1	0	1	0	0	0	0	0
8	0	2	0	0	0	0	0	0	1
9	0	0	0	0	0	0	0	0	0

FIGURE 6

NEXT ASSEMBLY QUANTITY TABLE

This is a concise mathematical representation of the information contained in the Assembly Parts Lists and this Table will form one of the building blocks of the mathematical theory.

go into A_2 is not a significant one and we could adopt any convenient system. Later, however, it becomes clear that using zeros makes the mathematics simple. In Figure 6, it can be seen that a row of zeros (e.g., the fourth row) indicates a top assembly; A_4 does not go into anything. A column of zeros indicates a detail part; thus, Article 6 does not require anything since it is not an assembly, but a part.

Let us stop for a moment now as we have reached in fact our first objective; the information contained in the Assembly Parts Lists has been put into an appropriate form for further mathematical discussion. Instead of Assembly Parts Lists, we will talk in terms of the Next Assembly Quantity Table as represented in Figure 6.

We proceed now to the determination of the parts requirements, that is, we develop a formula which tells us how many of each assembly and each part is required to meet any sales forecast.

THE TOTAL REQUIREMENT FACTOR TABLE

Before we discuss the determination of parts requirement we introduce some visual aids to clarify our concepts. Consider for this purpose Figure 7. It can be seen that A_1 is made up of one A_3 and two A_5's. A_2 is made up of two A_6's, one A_7, and two A_8's, etc. All this information is contained in the Next Assembly Table in a numerical fashion. Suppose now, we want to know all the articles that are required for each A_2.[4] The resultant diagram is shown in Fig-

[4] We mean by the statement, "articles required for each A_2," all the articles that go directly into A_2, *and* all the articles that go into these, *and so on.*

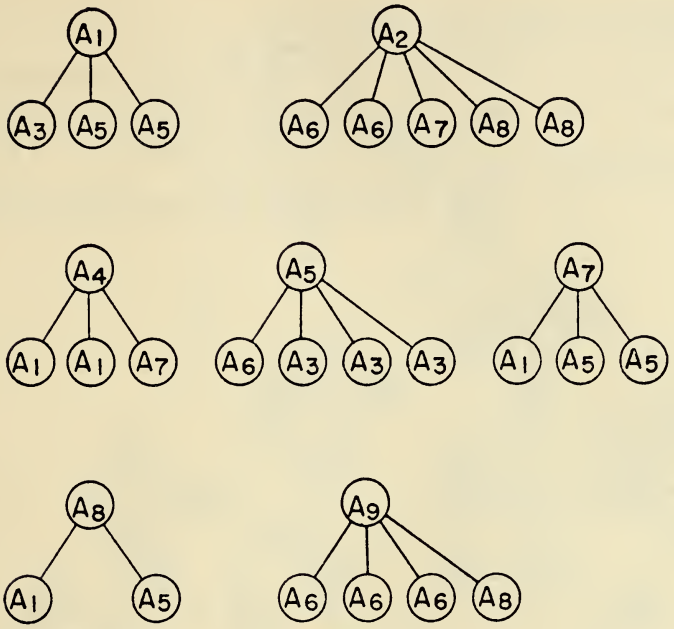


FIGURE 7

PICTORIAL REPRESENTATION OF NEXT ASSEMBLY QUANTITIES

It can be SEEN that A_1 is made up of one A_3 and two A_5's; A_2 is made up of two A_6's, one A_7, and two A_8's. It *cannot* be seen how many assemblies and parts are required *in total*, say, making up an A_2, when it is recognized that A_7 and A_8 are assemblies.

ure 8. This is not very convenient and so we present the same information in Figure 9 in a different form. Note that each article is shown only once. A_5 goes into A_7 twice so we have two arrows on the line going from A_5 to A_7. The insert shows that, say, A_5 goes into A_1, A_7, and A_8, twice, once, and twice, respectively. Figure 9 is a pictorial representation of certain columns of the Next Assembly Quantity Table; in order to fix our ideas it will be called the Gozinto Graph for A_2. Similar pictorial representations can be prepared for A_4 and A_9, and a composite Gozinto Graph for all our articles is shown in Figure 10. Figure 9 is much more simple than Figure 8 but contains the same information. However, from one point of view it is not quite so convenient. How many A_6's are required (in total) for each A_2? In Figure 8 one can answer this question by direct count of the A_6's. In Figure 9 some mental effort is required to answer the same question. However, this is the type of thinking that leads to the result we want as when dealing with thousands of assemblies we are not able to draw these various pictures. Suppose we wanted to figure out how many A_5's are required for each A_2. A_5 goes directly into A_1, A_7, and A_8 (see Figure 9), so we can make the statement that the

Total number of A_5's required for each A_2 =
[Number of A_5's going directly into each A_1]
× [Total number of A_1's required for each A_2]
+ [Number of A_5's going directly into each A_7]
× [Total number of A_7's required for each A_2]
+ [Number of A_5's going directly into each A_8]
× [Total number of A_8's required for each A_2]

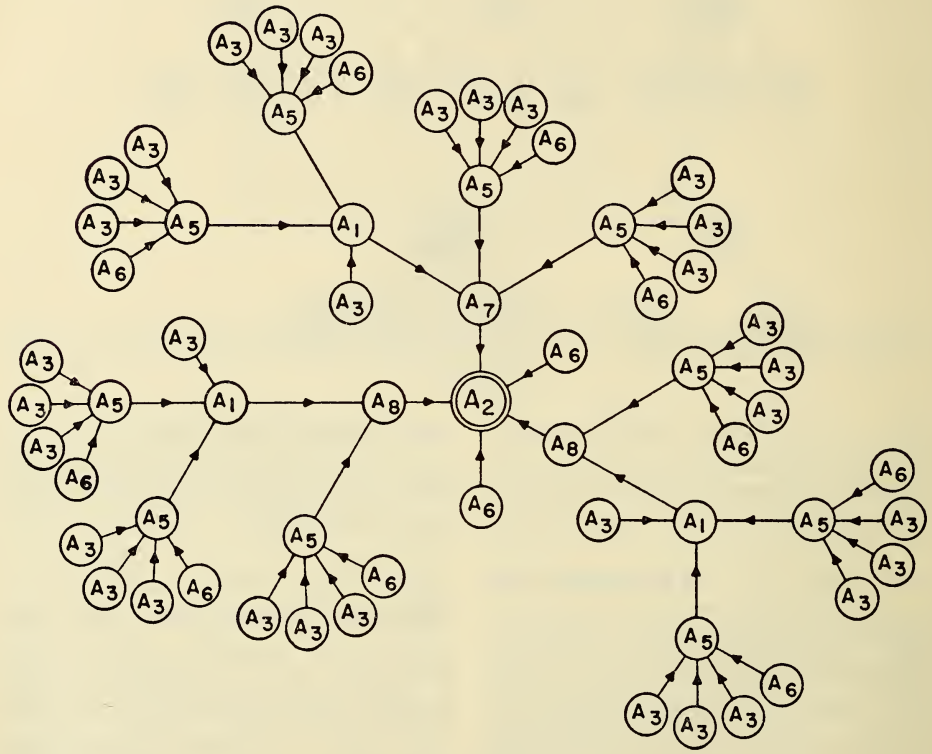


FIGURE 8

PICTORIAL REPRESENTATION OF TOTAL REQUIREMENTS

By a direct count it can be determined how many, say, A_3's are required for each A_2.

Note the important difference between these two statements—"Number of A_5's going directly into each A_8," and the statement "Total number of A_5's required for each A_2." The answer to the first statement is the Quantity 1, while the answer to the second one is the Quantity 3. In our Next Assembly Quantity Table we have the number of assemblies going directly into each other assembly listed, but we do not have the total number of assem-

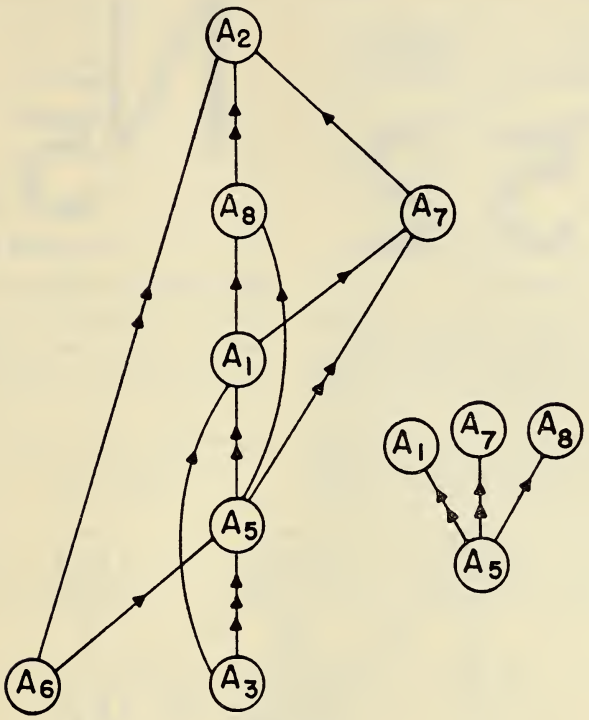


FIGURE 9

THE GOZINTO GRAPH FOR A_2

is a pictorial representation of the requirements for A_2. Each article is shown only once. The next assembly quantities are shown by the multiplicity of the arrows. Total requirements cannot be observed directly but can be deduced.

blies required for each other assembly.

Let us contemplate the above statement. It gives a relationship between total number of quantities required and next assembly quantities. It does not tell us how to compute the total number of quantities required from the next assembly quantities, as the various total numbers required and the next assembly quantities appear at both sides of the equation.

The same statement also implies some sort of a rule as we could also figure a relationship for how many A_5's we need, say, for each A_8. Very likely, this rule could be described in words; however, if one attempted to work out this rule, it would get lengthy and confusing. Clearly, what we need is a concise notation to describe the idea represented. And this is the point where mathematics comes in handy. Consider the diagram on the next page.

We have magnified the statement "Total number of A_5's required for each A_2." The way the picture was prepared suggests that instead of the long sentence, we could simply say $T_{5,2}$. Quite

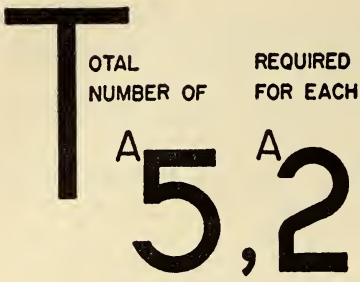


DIAGRAM I

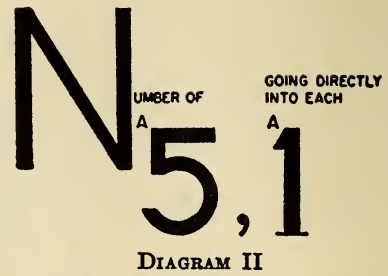


DIAGRAM II

similarly the diagram above suggests that instead of saying, "Number of A_5's going directly into each A_1," we should say $N_{5,1}$.

At this point bear in mind that the notation means nothing more or less than what we said. It is a concise statement of things we have known. However, we can now replace the statement in question with an equation:

$$T_{5,2} = N_{5,1} \cdot T_{1,2} + N_{5,7} \cdot T_{7,2} + N_{5,8} \cdot T_{8,2}$$

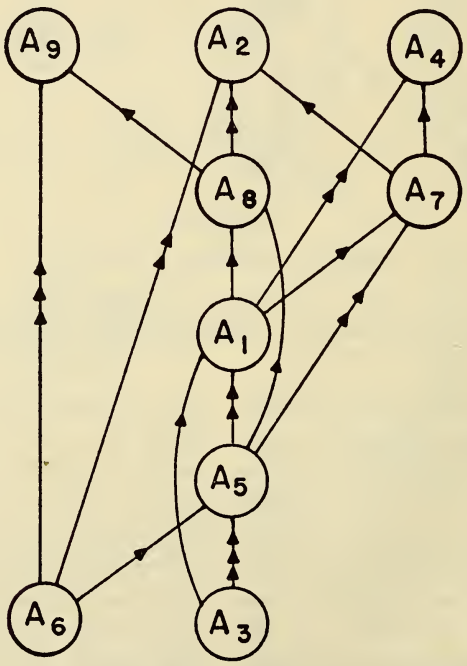


FIGURE 10

THE COMPOSITE GOZINTO GRAPH

is a pictorial representation of the parts requirements. The next assembly quantities can be observed directly by counting the arrows on each connecting line. Total requirements cannot be observed directly but can be deduced.

We have accomplished, then, our first objective—our specific statement is represented in a shorthand form.

Suppose for the moment that we know the answer we are attempting to find—that is, we have computed all the total requirements, which are all the various "*T*'s." They can be put in a table as shown in Figure 11. This table which we call the Total Requirement Factor Table, or briefly, the T Table, is very similar to the Next Assembly Quantity Table. Just as the latter one shows the "*N*'s," the new table shows the capital "*T*'s." One can see, for instance, that each A_7 requires one A_1, thirteen A_3's, nine A_5's, four A_6's, and one A_7.

Let us recognize that once the Total Requirement Factor Table is determined, our problem of answering the question of "how many" becomes very simple. Therefore, let us turn our attention to the general formulation of our equation, which formulation will lead directly to the determination of the Total Requirement Factor Table.

DETERMINATION OF THE "T" TABLE

Assuming for the moment that we have already computed all the *T*'s, let us focus our attention, say, on the second column of the T Table. Figure 12 shows the fifth row of the N Table and the second column of the T Table. Consider $T_{5,2}$, that is, the second element of the fifth row. We can say, as it is shown by our equation above, that

	1	2	3	4	5	6	7	8	9
1	1	3	0	3	0	0	1	1	1
2	0	1	0	0	0	0	0	0	0
3	7	33	1	27	3	0	13	10	10
4	0	0	0	1	0	0	0	0	0
5	2	10	0	8	1	0	4	3	3
6	2	12	0	8	1	1	4	3	6
7	0	1	0	1	0	0	1	0	0
8	0	2	0	0	0	0	0	1	1
9	0	0	0	0	0	0	0	0	1

FIGURE 11

TOTAL REQUIREMENT FACTOR TABLE

Observe say the *second* column relating to A_2. The *third* element from the top in this column relates to A_3 and displays the number 33. This means that thirty-three A_3's are required (in total) for each A_2. This can be confirmed by a direct count from Figure 8. (We have not explained yet how this above Table was computed.)

$T_{5,2}$ can be obtained by taking the left-hand side number from the N Table and multiplying it by the top number of the T Table; to this number we have to add the seventh number from the N Table multiplied by the seventh number from the top of the T Table; finally, we have to add the eighth number from the N Table multiplied by the eighth number of the T Table. This rule can be stated in a more simple form: Multiply the first number in the N Table with the first number in the T Table, multiply the second number with the second number, the third with the third, and so on, keeping in mind that we are combining a row with a column. This last rule is, of course, the same as the first, but it takes advantage of the zero's we have in the table.

Incidentally, to make this a good rule, we need both $N_{5,5}$ and $T_{2,2}$. The former was defined as zero and now we have to define $T_{2,2}$. We are at liberty to use any number, as the question of how many A_2's are required for each A_2 is not a significant one. However, it turns out that in order to make our rules uniform we have to assign to the diagonals of the T Table the value "1." It was precisely the same reason that we assigned the zeros to the diagonals of the N Table.

Mathematicians have developed a shorthand notation for sums of the kind we are discussing here. They simply write

$$T_{5,2} = \sum_k N_{5,k} T_{k,2}$$

using for summation the Greek capital letter Σ. This equation means exactly the same thing as the former one. The letter "k" indicates that the product should be computed for all values of "k."

It is quite clear that the equation that we have here works not only for $T_{5,2}$,

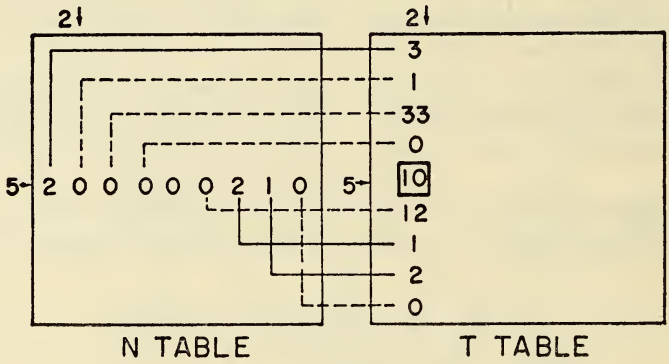


FIGURE 12

SCHEMATIC REPRESENTATION OF THE EQUATION $T_{5,2} = \sum_k N_{5,2}\ T_{k,2}$

The second number of the fifth row of the T Table equals the "scalar multiple" of the fifth row of the N Table and second column of the T Table:

$$10 = 2 \times 3 + 0 \times 1 + 0 \times 33 + 0 \times 0 + 0 \times 10 + 0 \times 12 + 2 \times 1 + 1 \times 2 + 0 \times 0$$

but for any element on the T Table. This can be written in mathematical form as

$$T_{i,j} = \sum_k N_{i,k} \cdot T_{k,j} \qquad i \neq j \quad (1)$$

The letters i and j simply mean that i can be any number and that j can be any number, though it is postulated that i must be different from j, as in that case $T_{i,i}$ takes the value "1."

Let us try this formula—consider, say, $i = 4$, $j = 2$; the T in question is $T_{4,2}$, the second element in the fourth row of the T Table. According to our rule, we have to combine the fourth row of the N Table (Figure 6) with the second column of the T Table (Figure 11). However, the fourth row of the N Table is a zero. Therefore, we can conclude that $T_{4,2}$ must be zero. This is not surprising as A_4 is a top assembly and A_2 does not require any A_4's.

Quite similarly we deduce from

$$T_{9,2} = \sum_k N_{9,k} \cdot T_{k,2}$$

that $T_{9,2}$ equals zero.

Let us continue now the computation of the rest of the elements in the second column of the T Table. We get

$$\begin{aligned} T_{8,2} &= \textstyle\sum_k N_{8,k} \cdot T_{k,2} \\ &= N_{8,2} \cdot T_{2,2} + N_{8,9} \cdot T_{9,2} \\ &= 2 \times 1 + 1 \times 0 = 2,^{5} \end{aligned}$$

$$\begin{aligned} T_{7,2} &= \textstyle\sum_k N_{7,k} \cdot T_{k,2} \\ &= N_{7,2} \cdot T_{2,2} + N_{7,4} \cdot T_{4,2} \\ &= 1 \times 1 + 1 \times 0 = 1 \end{aligned}$$

[5] Note that $T_{2,2} = 1$.

$$\begin{aligned} T_{1,2} &= \textstyle\sum_k N_{1,k} \cdot T_{k,2} \\ &= N_{1,4} \cdot T_{4,2} + N_{1,7} \cdot T_{7,2} \\ &\quad + N_{1,8} \cdot T_{8,2} \\ &= 2 \times 0 + 1 \times 1 + 1 \times 2 = 3 \end{aligned}$$

$$\begin{aligned} T_{5,2} &= \textstyle\sum_k N_{5,k} \cdot T_{k,2} \\ &= N_{5,1} \cdot T_{1,2} + N_{5,7} \cdot T_{7,2} \\ &\quad + N_{5,8} \cdot T_{8,2} \\ &= 2 \times 3 + 2 \times 1 + 1 \times 2 = 10 \end{aligned}$$

$$\begin{aligned} T_{6,2} &= \textstyle\sum_k N_{6,k} \cdot T_{k,2} \\ &= N_{6,2} \cdot T_{2,2} + N_{6,5} \cdot T_{5,2} \\ &\quad + N_{6,9} \cdot T_{9,2} \\ &= 2 \times 1 + 1 \times 10 + 3 \times 0 = 12 \end{aligned}$$

and finally

$$\begin{aligned} T_{3,2} &= \textstyle\sum_k N_{3,k} \cdot T_{k,2} \\ &= N_{3,1} \cdot T_{1,2} + N_{3,5} \cdot T_{5,2} \\ &= 1 \times 3 + 3 \times 10 = 33 \end{aligned}$$

We can see, therefore, that Equation 1 indeed allowed us to compute the second column of the T Table. Furthermore, it is clear that Equation 1 also can be used to compute all the other numbers on the T Table; therefore, we can conclude that Equation 1 does contain the necessary instruction for the determination of the T Table.

One more remark before we discuss our result in detail. It is to be pointed out that the computations had to be done in a very particular *sequence* since the T's appear on both sides of Equation 1. However, this need not be a cause for worry as one can try to compute the top element in the column; if this is not possible try the one below, proceed down to the bottom, and then again start on the top. This procedure eventually leads to all the numbers in the column. This method might not be the most efficient but it always works. A more careful study of the problem can

lead to a quicker procedure, but we shall not go into the details here.

THE MATHEMATICAL FORM OF A SALES FORECAST

We have said before that, once the Total Requirement Factor Table is determined, the problem of parts requirements is easy to solve. We propose to establish now the necessary mathematical formulism to determine the parts requirements.

In order to fix ideas, let it be supposed that twenty of A_2, thirty of A_4, eighty of A_9, and fifty of A_5 are specified by the sales forecast, and the problem is to determine how many A_5's are required. Clearly we have

Quantity of A_5's required =
(Total number of A_5's required for for each A_2) $\times$ (Sales forecast of A_2)
+ (Total number of A_5's required for each A_4) $\times$ (Sales forecast of A_4)
+ (Total number of A_5's required for each A_9) $\times$ (Sales forecast of A_9)
+ (Sales forecast of A_5)

We can put this statement into a mathematical formula. Let S_1, S_2, S_3, . . . etc., denote the sales forecast for articles A_1, A_2, A_3 . . . etc., and let the unknown requirements for A_5 be denoted by X_5, then

$$X_5 = T_{5,2} \cdot S_2 + T_{5,4} \cdot S_4 + T_{5,9} \cdot S_9 + S_5.$$

In our particular numerical case, we get

$$730 = 10 \times 20 + 8 \times 30 + 3 \times 80 + 50.$$

Using again the summation notation, the last equation can be written as

$$X_5 = \sum_k T_{5,k} \cdot S_k$$

where advantage is taken of the fact that $T_{5,5}$ equals 1. Finally, it is clear that a similar equation holds for any article, and so we write

$$X_i = \sum_k T_{i,k} \cdot S_k \qquad (2)$$

In actual practice many of the S's are zero as only some of the articles (say top assemblies and spares) are shippable.

In order to be sure that we understood our formula, let us try the case when $i = 3$. We get

$$X_3 = \sum_k T_{3,k} S_k = T_{3,2} S_2 + T_{3,4} S_4 + T_{3,5} S_5 + T_{3,9} S_9$$

and

$$2440 = 33 \times 20 + 27 \times 30 + 3 \times 50 + 10 \times 80,$$

showing that 2440 A_3's are required.

THE FINAL MATHEMATICAL FORMULATION

We have "solved" the problem of parts requirements in terms of the N and T "tables," our solution being expressed by Equations 1 and 2. As it happens, mathematicians have studied such "tables" in detail—they call them "matrices"—and a "Theory of Matrices" has

been developed. In Matrix Algebra, rules for the addition, subtraction, multiplication, and division of matrices are developed. We cannot go into the details of such a theory, but we will point out here that Equation 1 can be written as

$$[T] = [N][T] + [I]$$

or

$$[T] = \frac{[I]}{[I] - [N]}. \qquad (3)$$

where $[I]$ is the so-called "unit matrix." Equation 2 can be written as

$$[X] = [T] \times [S]. \qquad (4)$$

Finally, the two equations can be combined into a single equation

$$[X] = \frac{[I]}{[I] - [N]}\,[S] \qquad (5)$$

this being the final mathematical formulation of our problem. In order to bring into focus the mathematical methodology, we make now a complete statement using the language of a mathematician:

> Consider the manufacture of articles $A_1, A_2, \ldots$ and denote by $N_{i,j}$ the number of A_i's going directly into A_j. Let $S_1, S_2 \ldots$ denote the sales forecast and let $X_1, X_2, \ldots$ denote the (unknown) parts requirements. Then
>
> $$[X] = \frac{[I]}{[I] - [N]}\,[S]$$
>
> where $[N]$ is the (square) matrix formed by the $N_{i,j}$'s, $[I]$ is the unit matrix, $[S]$ and $[X]$ are the column matrices formed by the S's and X's.

CONCLUDING REMARKS

We have reached the end of our presentation—we have accomplished the transformation of our original verbal statement into a mathematical form. However, let us remind ourselves that we have studied only a very small part of the problem of production and inventory control and that the practical value of the theory lies in its extension to more complicated problems. As an example, we mention that our principal formula can be generalized to include the problems of scheduling. Instead of the sales forecast, $S_1, S_2, \ldots$, one must deal with the sales forecast *functions* $S_1(t), S_2(t), \ldots$, where each of these functions describes the sales forecast for individual planning periods such as days, weeks, months, etc. The requirements $X_1, X_2, \ldots$ are replaced by the requirement functions $X_1(t), X_2(t), \ldots$, designating the requirements for each planning period. The equations relating the $X(t)$'s to the $S(t)$'s become much more complicated as the effect of the various lead times, make spans, inventory policies, etc., all must be incorporated into the mathematics. A more detailed discussion of these problems lies beyond the scope of this paper and future publications are planned to report on this work. We conclude this paper now by elaborating on some of the advantages offered by the mathematical theory.

I. The mathematical statement deals with clear-cut and precise concepts. For instance, compare the exactness of the ideas involved in matrix multiplication with the vague thoughts represented by the word "explosion." The mathemat-

ical theory leads to a comprehension of matters that cannot be obtained by verbal discussions. A more and more complete mathematical statement of managerial problems and the mathematical solution of these problems will lead to an insight never heretofore realized.

II. A further outcome of the foregoing is the possibility of better ways of transmitting information on certain industrial practices. The writer has been constantly impressed by the great barriers that exist between various operating people and departments. The fact of the matter is that many systems and procedures are too complicated to be adequately described in words and, therefore, transmission of their description becomes very difficult. On the other hand, there is the likelihood that a mathematical formulation can be easily explained, as experienced by (in the rare instances of) lucid publications in scientific and engineering fields. The fact that current operating personnel are not trained along mathematical lines should not be considered an insurmountable obstacle.

III. Once the ideas are represented in mathematical form, specific managerial problems may be answered by manipulating the mathematical formulas with the aid of known mathematical techniques. These mathematical techniques need not be practiced by the operating personnel, as mathematicians well trained along these lines could be employed. It is to be emphasized that the mathematics used are not beyond the usual training of a mathematician with a Bachelor's or a Master's degree.

IV. In the introduction we touched upon the role that electronic computing machines will play in industry. However "intelligent" these machines will be, it still will be necessary to prepare the problems to be solved in a language digestible to the computing machine. Many current procedures are not completely formulated and are transmitted by verbal instructions or examples; clearly, this will not be sufficient for the electronic computing machine. Do we face the problem here of training our electronic computing machine experts in all ramifications of managerial tasks? A mathematical formulation gives great comfort in solving this problem. For instance, our mathematical formulation of the Parts Requirements problem can be given to a person trained in the use of electronic computing machines and he can solve the problem without going into the mass of details related to actual practices.

V. Finally, let us point out that such mathematical theories will lead to the general understanding of some managerial problems that so far have been treated only in an intuitive or haphazard fashion. Each result obtained in a mathematical theory will form a building block in the structure which eventually will form a discipline that truly could be called by the name "Management Sciences."

✦✦✦✦✦✦✦✦✦✦✦✦✦ PROTOTYPE MODELS IN OPERATIONS RESEARCH

RUSSELL L. ACKOFF

In the course of doing its work, Operations Research has repetitively encountered certain types of problems. As a result, Operations Researchers have come to define these types of problems in terms of their underlying structure. Concepts to describe these types of problems have been formulated, and tools have been developed, which are capable of solving them exactly or approximately. By way of a survey of the applications of Operations Research, I would like to review the various types of recurrent problems, indicate the forms in which they appear, and identify the scientific techniques and tools that have been developed for solving them.

Before discussing these problems, mention should be made of one essential aspect of the methodology of Operations Research. Operations Research's approach to executive-type problems consists of constructing a model of the system under study. A model is nothing more than a representation. Photographs, maps, and graphs are familiar types of models. Not all models, however, are useful in science. A scientific model is one which enables the researcher to manipulate the system modelled without affecting the system itself. By such manipulation the scientist can determine how the system can be changed so as to maximize its effectiveness. Consequently, the type of model that Operations Research uses is a type that represents the quantitative relationship between the effectiveness of the system and the various pertinent aspects (variables) of the system, some of which are subject to management control, and some of which are not. Then by mathematical manipulation the Operations Researcher attempts to find the values of the controllable variables for which the system's effectiveness is maximized.

The general form of an Operations Research model, then, is

$$E = f(x_i, y_j)$$

where E is the measure of the system's effectiveness (or ineffectiveness), x_i is the set of controllable variables, and y_j is the set of uncontrollable variables. For example, in a production problem E may be the total cost of production, one of the controllable variables may be the size of the production, and one of the uncontrollable variables may be customers' demand. The Operations Research problem is to find those values of the variables represented by x_i

Reprinted with permission from "A Survey of Applications of Operations Research" in *Proceedings of the Conference on Case Studies in Operations Research*, Case Institute of Technology, 1956, 9–17.

which will maximize effectiveness E (or minimize ineffectiveness).

The seven classes of recurrent problems to be considered are inventory, allocation, waiting-time, routing, replacement and maintenance, search, and competitive.

INVENTORY PROBLEMS

Shortage costs, which involve the costs of delay or the inability to satisfy a customer, increase as inventory decreases. Thus, as inventory carrying costs decrease shortage costs increase. If we try to minimize the costs associated with hiring, firing, and training personnel, we must produce for inventory during slack period. Inventory problems, then, involve the attempt to obtain a minimum sum of these conflicting types of cost. Specifically, inventory problems may appear in three different forms.

In the first, a question arises when the time of production is fixed but the quantity that is to be produced is variable. The problem here is to decide how much should be produced. For example, on an assembly line devoted to one product the problem is not when to produce, but how much to produce.

In the second, production quantities are fixed; for example, in a chemical process in which products are made in batches, the question is, when and how frequently to produce or purchase.

A third or combined type of inventory problem occurs when we can control both the amount produced and the frequency of production. This problem was solved mathematically for very restricted conditions as early as 1915, by what is now known as an "economic lot size equation." Since 1915, and particularly in the last few years, techniques have been developed for solving this problem where demand is not known but is predicted, where production or purchasing lead-time is uncertain, where the quantity delivered fluctuates around the quantity ordered, and where all three of these uncertainties operate simultaneously. The problem can also be solved for the purchasing situation in which quantity discounts must be taken into account.

It might be helpful to cite some instances in which the three types of inventory problems have been solved. First, consider the case where the production time is fixed and the production quantity is to be determined. Such a problem has been solved for the production of a pharmaceutical which is produced by a fermentation process. The process is such that fermenters must be set every day. The problem was to determine how many fermenters should be set each day. It was not quite this simple, however, because the process yielded three different forms of the product depending on how the yield of the fermenters were further processed. Consequently, the solution required specification of not only the number of fermenters to be set each day but the amount of each form of finished product that was to be produced.

The second type of inventory problem involves fixed production or purchasing quantities, but controllable time of acquisition. Such a problem was recently solved for a large public utility company which furnishes electricity to one of our major urban areas. The problem consisted of determining when additional generating capacity

should be added to the system. This may not appear to be an inventory problem, but it is. If the additional capacity is added too early there is an associated cost of carrying an unused investment until the capacity is required. This is the equivalent of an inventory-carrying cost. On the other hand, if capacity is added too late, shortages of current will occur. Such shortages affect the probabilities of obtaining price adjustments from the public utilities commission. Hence, there was a very real shortage cost associated with the problem.

The third type of inventory problem involves determining both the timing and the quantity produced or purchased. This problem arose recently in a peculiar garb for an airline which wanted to determine how frequently it should conduct classes for training stewardesses and how large the classes should be. The stewardess school is analogous to a production facility turning out a product. There was a set-up cost associated with running a class, an inventory holding cost which consisted of paying unused stewardesses in the system, and a shortage cost associated with cancellation of flights or the emergency action taken to prevent such cancellation.

This type of problem has also been solved in a variety of jobshop operations, particularly in metal working industries. For example, it has been applied to the production of turret lathe parts, diesel engines, textile manufacturing equipment, and automotive parts.

ALLOCATION PROBLEMS

The second class of problems that we will consider consists of allocation problems. Such problems are also of three types, the first of which is defined by the following conditions:

(1) There are specified jobs to be done or activities to be performed.

(2) There are specified facilities or resources available.

(3) There are alternative ways of performing the activities or doing at least some of the jobs, with the available resources or facilities, some of which ways are more efficient than others.

In other words it is not possible to do every job in the best possible way. The problem is to determine how to do each job so as to maximize the overall effectiveness (e.g., minimize the total cost).

The second type of allocation problem occurs when the jobs are specified but the facilities are subject to control. Hence, the problem is to determine what set of facilities will yield maximum effectiveness, taking into account the cost of the facilities as well as the cost of operating them.

The third type of allocation problem arises when the facilities or resources are specified and the jobs or activities are to be specified.

There are many illustrations of the solution of the first type of problem (both jobs and facilities specified). For example, in one steel company all production is to order and is planned each month. On the average there are six possible ways of manufacturing the product specified on each order. That is, different equipment can be used to do the same jobs. One of these different production practices is usually better than the others. But in the experience of the company there had never been

a mixture of orders to fill in any one month which allowed each order to be filled in the best way. If they tried to do so, some facilities would have been overloaded and others underloaded. The solution of the problem consisted of finding for (each month) that combination of production practices which would allow every order to be filled in such a way as to minimize the total cost of production.

Among other examples of this type of problem which have been solved are the following. How much and what raw materials should be shipped from five smelting plants to fifteen fabricating plants so as to minimize the sum of the transportation cost? How should the time of salesmen be allocated to accounts of the General Electric Lamp Division so as to maximize the volume of orders received?

The second type of allocation problem involves specified jobs but controllable resources. Some examples of this type of problem which has been solved are the following. How many warehouses in what locations will minimize the costs of distribution to a specified set of customers? For an airline how many crew bases in what locations will minimize the sum of the away-from-home costs and the base operating costs?

The third type of allocation problem involves specified facilities but controllable jobs. For example, the Standard Oil Company of New Jersey has determined for one of its refineries what mix of product in what quantities will maximize the profits realized from the refinery's output. In another case, one of a set of alternative new products was selected so as to minimize the total cost of production of new and old products with the available production facilities.

Allocation problems are solvable by a variety of techniques, but by far the best known and most frequently used one is Linear Programming. (Non-Linear Programming methods are also available and have also been used.) The development of this technique and of procedures for "putting it" on high speed electronic computers has made it possible to solve quickly even large allocation problems.

WAITING-TIME PROBLEMS

These problems arise in the following type of situation. Units requiring service (called "customers" even if they are inanimate) arrive at a service point at which facilities are available for rendering service. The customers and the facilities are seldom in perfect balance; that is, one or the other is kept waiting at various times. Costs are associated with waiting customers (e.g., loss of business, wages for idle employees, or in-process inventory cost) and also with idle facilities.

There are four major aspects of this servicing process any one or combination of which may be subject to control: the number of service facilities, the arrival rate of customers, the order in which customers are served, and the arrangement of the service facilities where more than one service operation is performed.

Only four of the many possible types of waiting-time problems have been explored to date. These are the facilities problem, the scheduling problem, the

sequencing problem, and the line-balancing problem.

The facilities problem arises when (1) the arrivals of customers are not subject to control, (2) the ordering of available customers for service is specified by a rule, (3) the number of service facilities is subject to control, and (4) each facility can perform all the required service. For example, a British Operations Research team recently determined what number of ore unloading cranes should be provided at a harbor facility so as to minimize the sum of the costs arising from waiting ships and idle ore-handling equipment. At Boeing Aircraft a study determined the number of supply room clerks required to minimize the sum of the costs arising from idle clerks and workers who came for supplies. The Port of New York Authority determined the optimum number of toll booths to operate at its various bridges and tunnels at different times of the day.

The scheduling problem is similar to the facilities problem except that the timing of arrivals rather than the amount of facilities, is subject to control. It arises, for example, in scheduling transportation facilities into an airfield, a train terminal, or an unloading dock.

The scheduling and the facilities problems are both solvable by use of the Monte Carlo Method which is a procedure for simulating the process under study. By such simulating various alternative procedures can be tried, and one can converge on the best solution. There is also a mathematical theory, called Queuing Theory, which can be applied to a restricted number of these problems. The Monte Carlo Method can also be used to solve waiting-time problems where the order of service is subject to control and where the service is provided by a chain of service facilities (e.g., the various servers at a cafeteria counter).

The sequencing problem arises when (1) a set of customers are available for service or their availability can be controlled, (2) the service required must be performed by a chain of specified facilities, (3) the sequences of operations that must be performed for the various customers differ either in the operations required, the order of operations, and/or the duration of the operations. The problem is to determine the sequence (order) in which the customers should be serviced (i.e., jobs should be done) so as to minimize the total elapsed service time.

Using a high speed electronic computer and Monte Carlo procedures, a group in the General Electric Company developed a method for sequencing production of an entire factory. This method did not yield an optimum sequence but it did provide a good approximation. In this problem all the uncertainties associated with absenteeism, machine breakdowns, and variable quality and quantity of machine output were taken into account.

Just recently mathematical theory has been developed for finding optimal sequences for a very restricted class of sequencing problems. We can, however, expect large and rapid advances to occur in this area in the near future.

The line-balancing problem arises when (1) arrivals are subject to control, and (2) all customers require the same

multiple operations to be performed on them. The problem is how to combine the operations to be performed into work stations so as to minimize the total service time. Ideally, the solution consists of forming the work stations so that the service time at each is equal. Then the customers can flow at a constant rate and not accumulate before any station.

This type of problem arises in the design of production and assembly lines. Several techniques for obtaining approximately optimal solutions to this problem have recently been developed. One was applied to the design of an assembly line in one of the General Electric Company's large appliance plants.

ROUTING PROBLEMS

There is a classical mathematical problem called the traveling-salesman problem. In this problem a salesman has to visit each of a number of cities once and only once and return to his point of origin. He must do this in such a way as to minimize the travel cost or travel time. The problem, then, is to find an optimal route. When the number of cities is very small we can enumerate the possibilities and evaluate each of them. But if there are a large number of cities the arithmetical approach becomes prohibitive. Although no general mathematical short-cut exists for solving this problem, solutions can sometimes be obtained by the "Assignment Method" of Linear Programming. In addition, methods are available which generally yield economical results although they are not optimal.

Of what significance is this mathematical problem? Recently it was found that a common industrial production problem has the same structure. Consider a production or assembly line on which a variety of products is made. Some of these products have common parts. Hence the setup cost for a production run of any one product depends on which product preceded it. The problem is to find that sequence of the production runs which minimizes the sum of the setup costs. This problem has exactly the same structure as the traveling-salesman problem.

This problem was recently attacked by the Mullens Manufacturing Company with respect to their kitchen sink production line. A procedure for obtaining an approximation to an optimal solution was developed which yielded large reductions in total setup costs.

REPLACEMENT AND MAINTENANCE PROBLEMS

Replacement and maintenance problems are the same in structure. They differ only in the perspective of the decision-maker. Maintenance problems involve the replacement of components of a unit. For example, replacement of tires is part of automobile maintenance. Replacing light bulbs, paint, or window panes is part of factory maintenance.

Replacement problems may be divided into two general classes depending on the nature of the equipment involved. The first class involves equipment whose efficiency decreases with use or the passage of time; for example, turret lathes, trucks, or bulldozers. A

replacement decision for such equipment involves two conflicting types of cost: that arising from the inefficiency of the equipment and that arising from the cost of new equipment. The problem is to determine the times at which replacement results in a minimum sum of these two costs.

Although no general solution to this class of problems is available, a number of special solutions have been found. These specialized techniques were developed by industrial engineers and economists and have been used widely. They have been applied to a variety of problems involving industrial equipment and transportation facilities. More recently new techniques have been developed in Operations Research for solving this type of problem. To cite but one instance, the technique of Dynamic Programming has been shown to be applicable in this area. Operations Research projects have involved finding policies of replacement of rail in railroad systems, trucks in a truck fleet, and aircraft in military units.

The second type of replacement problem involves units which do not degenerate significantly with use but which, after various amounts of usage, fail or die. Such units are electric light bulbs, electron tubes, etc. This type of problem involves three costs: The cost of the equipment itself (which is minimized if each unit is used until it fails), the cost of failure (e.g., work stoppage which is maximized if each unit is used until it fails), and the cost of replacement (which is usually higher per unit for individual replacement than for group replacement). The problem is to determine the time (if any) to make a group replacement and which units should be replaced.

Actuarians and mathematicians have worked on various phases of this problem for many years. Out of this work has come a mathematical theory which can be used to find very close approximations to optimal replacement policies for simpler replacement problems. More complex problems can be fruitfully solved (with a high degree of approximation to the optimum) by Monte Carlo techniques.

These methods have been used to develop rules for the replacement of light bulbs. They have also been used to regulate replacement of electron tubes in airborne radar and tires in a truck fleet. Policies for maintenance of industrial equipment have also been established by use of replacement theory and Monte Carlo procedures.

SEARCH PROBLEMS

A search problem is characterized by the need for designing a procedure to collect information on the basis of which one or more decisions are made. Such decisions can be in error because of errors in the information. Errors in the information, in turn, can arise from either (a) the method of observing and/or (b) the number of observations. The first of these are errors of observation, and the second are sampling errors. Relative to a fixed amount of money, time, or other resources to be used in collecting information, if more careful observations are made (so as to decrease the errors of observation), fewer observations can be made (thereby increasing sampling error).

Search problems can be classified as follows: (1) Relative to fixed resources, what and how many observations should be made (and with what amount of care) so as to minimize the expected cost of errors in the dependent decisions? For example, one railroad did not have enough time available to examine every record of the division of revenues for carload shipments in which it is the initiating or intermediate carrier. (The terminating carrier apportions the revenue it collects.) Errors in auditing depend on how much time is spent on a record, but the total amount of errors found also depends on the number of records examined. An optimal auditing procedure was found for this railroad. This same type of problem arose during the war when it was necessary to determine how patrol planes should be used to search for submarines.

(2) How much time or money should be spent in the search operation and how should it be used so as to maximize the difference between the expected gain from use of the information and the loss incurred collecting it. For example, the M. A. Hanna Company was confronted with the problem of determining how many exploratory holes it should drill in an ore field and where it should drill them, so as to obtain information on the basis of which the company could decide whether or not to mine the area.

(3) In both preceding cases the problem was to design a search procedure to find things the locations of which were not known. In a third type of search problem we can control the location of the items or information sought but cannot control the search procedure. The problem is how to locate the items so as to maximize the effectiveness of the search procedure. For example, in placing items in a super market or department store there is the problem of locating these items so as to maximize the chances that the customer (whose search procedure is not under control) will find what he wants.

There are a number of scientific techniques which are useful in solving search problems. The theory of probability sampling, however, is perhaps the most important of these. This is a powerful technique whose potentiality is only beginning to be explored by business and industry.

COMPETITIVE PROBLEMS

A competitive problem is one in which the efficiency of the decision maker is affected by the decision(s) of others. For example, the effectiveness of a company's advertising and pricing policies depends on the advertising and pricing policies of its competitors.

The game is, perhaps, the most generally familiar type of competition situation. The Theory of Games, which is directed toward obtaining optimal solutions of games, has developed at a considerable rate over the last ten years. But as yet solutions are available for only very simple games and, consequently, few (if any) direct applications of the theory have been made to practical problems. The theory has considerable impact within science, par-

ticularly in the area called "decision theory." Despite the fact that it is not directly applicable to practical industrial problems it can have a significant effect on the way we go about solving practical competition problems. Williams has noted this effect as follows:

> While there are specific applications today, despite the current limitation of the theory, perhaps its greatest contribution so far has been an intangible one: the general orientation given to people who are faced with overcomplex problems. Even though these problems are not strictly solvable—certainly at the moment and probably for the indefinite future—it helps to have a framework in which to work them. The concept of a strategy, the distinctions among players, the role of chance events, the notion of matrix representations of the payoffs, the concepts of pure and mixed strategies, and so on, give valuable orientation to persons who think about complicated conflict solvations. (J. D. Williams, *The Compleat Strategyst,* McGraw-Hill Book Co., Inc., New York, 1954, page 27.)

It is worth citing an indirect impact of Game Theory on the study of military strategies because its implication for the industrial situation are beginning to be explored. This impact takes place through the derivation of what might be called the "conservative competitive principle." This principle directs a decision maker to establish a new competitive policy if and only if he can show that it will net him a gain regardless of what his competition (the enemy) does. On military research projects devoted to the development of new weapon systems, for example, it is not unusual to have a separate team set up to represent the enemy. The situation is set up to be as favorable to "the enemy team" as possible. Under these circumstances an "operational game" is played. Such a procedure provides a basis for maximizing the minimum expected gain in a competitive situation.

Another type of competitive situation is that in which bidding for property or the right to render or receive service is involved. The concepts and techniques of decision theory have recently been found to be applicable to such problems. A few successful applications have already been made. However, industrial security has prevented publication of details.

MIXED PROBLEMS

A problem as it arises in a real industrial situation can seldom be characterized completely by placing it in one of the seven classes of problems we have considered. Each problem type we have considered is in a sense an abstraction from reality. Reality is seldom, if ever, as simple in structure as are these types of problem. For example, complete control of a production process usually involves at least (1) determining production quantities (an inventory problem), (2) machine loading (an allocation problem), and (3) sequencing production runs. Or again, associated with any policy for replacing items that fail we also have an inventory problem associated with the replacement parts or units. To cite one last example, an optimum schedule for

an airline cannot be made without simultaneously considering what the plant-maintenance policy should be.

As yet we have not developed procedures for deriving exact optimal solutions where several problem types are combined, but we do have ways available for approximating such solutions. The development of optimal solutions for these more generalized problems will be the next major stage in the development of Operations Research.

part ✦ 3

The Methodology of Operations Research: Techniques

"Neither a surgeon nor an OR man is made by a bag of tools. The tools are means. The OR man must be good first of all at finding out what needs to be done. Then he can select his tools, or maybe make a new one." [1] Awareness of this fact—that OR is not merely a collection of tools whose mastery makes one, *ipso facto,* a competent analyst—is an important initial step toward an understanding of the discipline. However, the methods used by workers in any discipline are a significant, identifying characteristic of the discipline and this is especially true of OR. As a prominent British operations analyst has said, ". . . what constitutes operational analysis is not knowledge but know-how. Problem solving technique is the bond that links operational research workers in diverse fields." [2] Although this statement might be disputed by other operations analysts, it contains sufficient truth to warrant the conclusion that a grasp of the nature of OR requires some acquaintance with the techniques employed by workers in the discipline.

In fact, it is really very important for executives to have more than a mere acquaintance with the tools of operations researchers. To be sure, no executive can hope to achieve expert mastery of these tools unless he has had or can obtain considerable training in mathematics. Such mastery is not, however, what the executive needs. It is no more necessary for the manager to be an operations

[1] Solow, Herbert, "Operations Research Is Methods: What Are They?," *Operational Research* in Business," *Fortune,* February, 1956, 148.

[2] Jessop, W. N., "Operational Research *Quarterly,* June 1956, 7:2, 50.

analyst than it is for him to be an accountant or lawyer. The services of experts in all these areas can be hired. However, the executive must be able to assist, supervise, and evaluate the work of the expert. And to do this when employing an operations researcher, the executive needs to understand the nature and purposes of the analyst's techniques even if not their complete detail.

What the executive needs is sufficient depth of understanding of techniques to enable him to assist the operations analyst in the selection of problems for investigation, in the definition of the critical factors in a problem as well as the relationships among these factors, and in the interpretation of the results of the investigation. What he needs is a clear recognition of the capabilities and limitations of the analyst's techniques so that he can judge whether they are applicable to the problems for which they are being used, and whether their use takes into account the important elements of these problems. "In the last analysis," as one writer has noted, "no techniques, however sophisticated and powerful, can replace the judgment, insight and wisdom of the sophisticated executive. Engineers and doctors have long known this. But the best engineer and the best doctor will be acquainted with the latest, most sophisticated methods which can help him make the most effective use of this judgment." [3]

To comprehend the nature, purposes, capabilities, and limitations of the tools of OR, however, requires some knowledge of their components. It requires a familiarity with their fundamental concepts and their essential structure as well as with the types of operations involved in their use. The third section of this book, therefore, provides a description of some of the more important techniques used by operations analysts and the types of problems for which they are used.

The difficulties encountered by executives in making decisions seem to have two principal sources—complexity and uncertainty. Often the executive's problem involves an enormous number of variables interrelated in many different ways so that hundreds and even thousands of possible solutions exist. Faced with such a problem, the executive usually finds that he cannot even define, much less evaluate all of the possible solutions. He cannot spend the time that this requires and, beyond that, he just cannot take into account so many variables and so many relationships, simultaneously, in his mental calculations. As a result, there is always the danger that he has overlooked or failed to evaluate properly some solution for his problem that is "best."

In addition, the executive's problem may involve random or chance variables. The values which such variables assume cannot be predicted exactly. They can be predicted only in the terms used to forecast the outcome of a spin of a roulette wheel. That is, they can be predicted only in probabilistic terms. Sales of a product during a particular week, the outcome of a salesman's call, and the time it will take to process a customer at the checkout station of a supermarket are ex-

[3] ———, "A Guide to Operations Research Methods," *Cost & Profit Outlook*, Alderson Associates, Inc., April 1957, **10**:4, 1.

amples of such variables which are important in business. When random variables are important in a problem, the executive usually finds that he is unable to predict the consequences of any solution with satisfying confidence. He is frequently, in fact, so uncertain about the outcomes of each of his possible solutions that he is unable to determine which of them should be adopted.

There is another factor, in addition to chance variables, which makes prediction of the consequences of possible solutions extremely difficult. This is the existence of competitors. Clearly, the outcome of any contemplated solution to many business problems depends upon the actions of rivals. Yet it is usually impossible to forecast precisely what these rivals will do. The result is that the executive is unable to evaluate his alternative solutions. Again, he cannot estimate their outcomes with desirable confidence.

Like the actions of competitors, other events over which executives have no control produce uncertainty. Whether there will be a recession or a boom, whether there will be a hot war, cold war, or rapid and total disarmament, and whether a particular foreign exchange rate will rise or fall are examples of such events. From the point of view of the individual executive they are tantamount to "acts of God." Nevertheless, for many business problems the "state of nature" has a considerable bearing on the effectiveness of each solution that may be considered. And since the executive cannot know which "state of nature" will characterize the relevant future, he often cannot determine his "best" course of action.

Finally, uncertainty about the consequences of action alternatives may originate from an entirely different source. It may arise because records of past experience with similar or analogous problems are inadequate or simply not available. This may be because adequate records have never been kept or because the situation being faced is different enough from any in the past to make it truly unique. Where either of these conditions prevails and there are no records which could provide information on which the executive could base his predictions, the executive is likely to be completely adrift in a "sea of uncertainty."

Thus sheer complexity, variability arising from the presence of chance factors, competitors, "acts of God," and lack of information are the chief obstacles to decision making of high quality in business. The tools used by operations analysts can be looked upon as means for coping with one or more of these obstacles. It is, in fact, possible to group the tools of OR on the basis of the particular obstacle with which they can best deal. Such a classification is not only convenient but may actually be illuminating, and so it has been used in this section. The first segment, therefore, describes mathematical programming, dynamic programming, symbolic logic, and factor analysis—all of which are methods for handling complexity. The second segment, on the other hand, contains discussions of probability theory, queuing or waiting-line theory, the theory of games and decision theory—techniques which are useful in dealing with problems involving variables whose values or states are not precisely predictable. Finally, the third segment contains discussions of sampling theory, statistical inference, the Monte Carlo

method and simulation—methods used by operations analysts when needed information is inadequate or unavailable.

Before proceeding to the readings themselves, several caveats are in order. First, the classification of OR tools which has been used is open to criticism. Some of the techniques which have been placed in a particular group could, with equally compelling reason, have been placed in another. The fact is that some of the tools are multi-purpose and it is impossible to determine in such cases which purpose is dominant. Secondly, most business problems cannot be dealt with fruitfully through the use of a single technique. Often these problems are not only complex but involve both random variables and lack of critical information as well. To analyze such problems, a combination of tools drawn from all these categories is required. Finally, not all of the techniques described in this section are of proven usefulness. Some are, of course. Others have had only limited application to this time, and so, although they appear to have considerable potential, they cannot, as yet, be regarded as proven useful. Still others seem to have only limited possibilities of practical application but they have been included because they employ concepts which can contribute to improved problem solving.

A Tools for Coping with Complexity

I MATHEMATICAL PROGRAMMING

MATHEMATICAL PROGRAMMING: *better information for better decision making*

ALEXANDER HENDERSON AND ROBERT SCHLAIFER

In recent years mathematicians have worked out a number of new procedures which make it possible for management to solve a wide variety of important company problems much faster, more easily, and more accurately than ever before. These procedures have sometimes been called "linear programing." Actually, linear programing describes only one group of them; "mathematical programing" is a more suitable title.

Mathematical programing is not just an improved way of getting certain jobs done. It is in every sense a *new* way. It is new in the sense that double-entry bookkeeping was new in the Middle

Harvard Business Review, *May–June 1954*, 32:3.

AUTHORS' NOTE: The authors wish to express their gratitude to Charles A. Bliss, W. W. Cooper, and Abraham Charnes for their invaluable assistance in the preparation of this article.

Ages, or that mechanization in the office was new earlier in this century, or that automation in the plant is new today. Because mathematical programing is so new, the gap between the scientist and the businessman—between the researcher and the user—has not yet been bridged. Mathematical programing has made the news, but few businessmen really understand how it can be of use in their own companies.

This article is an attempt to define mathematical programing for businessmen, describe what it means in practice, and show exactly how to use it to solve company problems. We have divided the article into four sections:

Part I is addressed specifically to the top executive. Here are the salient points about mathematical programing which the man who makes company policy needs to know.

Part II is addressed to executives directly responsible for the organization and administration of operations where mathematical programing could be used and to the specialists who actually work out the problems. This part is based largely on case examples which are typical of the kinds of problems that can be handled.

Part III shows management how to use mathematical programing as a valuable planning tool. In many situations programing is the only practical way of obtaining certain cost and profit information that is essential in developing marketing policy, balancing productive equipment, making investment plans, and working out rational decisions on many other kinds of short-run and long-run problems.

In addition, to be used in connection with Part II, there is an appendix providing actual instructions for working through the most frequently useful, quick procedure for solving a common class of business problems.

PART I. BASIC PRINCIPLES

Production men usually have very little trouble in choosing which machine tool to use for a given operation when there is free time available on every tool in the plant. Traffic managers usually have little trouble in choosing which shipping route to use when they are able to supply each of their customers from the company's nearest plant. The manager of a refinery usually has little trouble in deciding what products to make when he has so much idle capacity that he can make all he can sell and more.

Except in the depths of depression, however, the problems facing management are usually not this simple. Any decision regarding any one problem affects not only that problem but many others as well. If an operation is assigned to the most suitable machine tool, some other operation on some other part will have to be performed on some other, less suitable tool. If Customer A is supplied from the nearest plant, that plant will not have sufficient capacity to supply Customer B, who also is closer to that plant than to any other. If the refinery manager makes all the 80-octane gasoline he can sell, he will not have capacity to satisfy the demand for 90-octane gasoline.

BUSINESS PROGRAMS

The general nature of all these problems is the same. *A group of limited re-*

sources must be shared among a number of competing demands, and all decisions are "interlocking" because they all have to be made under the common set of fixed limits. In part, the limits are set by machine-tool capacity, plant capacity, raw materials, storage space, working capital, or any of the innumerable hard facts which prevent management from doing exactly as it pleases. In part, they are set by policies established by management itself.

When there are only a few possible courses of action—for example, when a company with only two plants wants to supply three or four customers at the lowest possible freight cost—any competent scheduler can quickly find the right answer. However, when the number of variables becomes larger—when a company has a dozen factories and 200 or 300 customers scattered all over the country—the man with the job of finding the best shipping pattern may well spend many days only to end up with a frustrated feeling; though he thinks he is close to the right answer, he is not at all sure that he has it. What is worse, he does not even know how far off he is, or whether it is worth spending still more time trying to improve his schedule. The production manager who has 20 or 30 different products to put through a machine shop containing 40 or 50 different machine tools may well give up as soon as he has found *any* schedule that will get out the required production, without even worrying whether some other schedule would get out the same product at a lower cost.

Under these conditions business may incur serious unnecessary costs because the best program is not discovered. Another kind of cost is often even more serious. The few direct tests which have been made so far show that intelligent and experienced men on the job often (though by no means always) come very close to the "best possible" solution of problems of this sort. But since problems of such complexity can almost never be handled by clerical personnel, even these good cut-and-try solutions are unsatisfactory because they take up a substantial amount of the time of supervisory employees or even of executives.

The time of such men is the one thing that management cannot readily buy on the market. If it is all used up just in getting the necessary information, there is nothing left for the next step, making sound decisions. Often this produces a sort of inertia against *any* change in the status quo; it is so hard to find out the cost or profit implications of a proposed change or series of changes that management simply gives up and lets existing schedules and programs stand unchanged. Conversely, if better information were available more easily, management would be less tempted to drop important questions without investigation or could make better decisions as a result of investigation.

A Routine Procedure Many of these complex and time-consuming problems can in fact be solved today by mathematical programing. The purely routine procedures of which it is comprised can be safely entrusted to clerical personnel or to a mechanical computer. Such procedures have already been success-

fully applied to practical business problems, some of which will be described in the course of this article.

The word mathematical may be misleading. Actually the procedures go about solving problems in much the same way as the experienced man on the job. When such a man is faced by a problem with many interlocking aspects, he usually starts by finding a program that meets the minimum requirements regardless of cost or profit, and then tries out, one by one, various changes in this program that may reduce the cost or increase the profit. His skill and experience are required for two reasons: (a) to perceive the desirable changes and (b) to follow through the repercussions of a single change on all parts of the program.

What "mathematical" programing does is to reduce the whole procedure to a simple, definite routine. There is a rule for finding a program to start with, there is a rule for finding the successive changes that will increase the profits or lower the costs, and there is a rule for following through all the repercussions of each change. What is more, it is *absolutely certain* that if these rules are followed, they will lead to the best possible program; and it will be perfectly clear when the best possible program has been found. It is because the procedure follows definite rules that it can be taught to clerical personnel or handed over to automatic computers.

COST INFORMATION

Quick and inexpensive calculation of the best possible programs or schedules under a particular set of circumstances is not the only benefit which management can obtain from this technique. The same complex situation which makes it difficult to find the best possible schedule for the entire operation makes it difficult to get useful cost information concerning details of the operation. When every operation in the shop can be performed on the most suitable machine tool, the cost of any particular operation can be obtained by the usual methods of cost accounting. But if capacity is short, then the true cost of using a machine for one particular operation depends in a very real sense on the excess costs incurred because some *other* part has to be put on a less suitable machine. To illustrate further:

> If the production of 80-octane gasoline is carried to a point where less 90-octane can be produced than can be sold, the profits which failed to be made on 90-octane must certainly be kept in mind when looking at the stated profits on the 80-octane.
>
> When a company is supplying some (but not all) of its eastern customers by bringing in supplies from the West Coast, additional cost will be incurred by giving one of these customers quick delivery from a nearby plant, even though the actual freight rate from the nearby plant is lower than the rate from the West Coast.

Any time that the programing procedure will solve the basic problem of determining the most profitable over-all schedule, it will also produce *usable* cost information on parts of the whole operation. In many cases this information may be even more valuable than the basic schedule. It can help manage-

ment decide where to expand plant capacity, where to push sales, and where to expend less effort, or what sorts of machine tools to buy on a limited capital budget. In the long run, sound decisions in matters of this sort will pay off much more substantially than the choice of the best shipping program in a single season or the best assignment of machine tools for a single month's production.

LIMITATIONS

Mathematical programing is not a patented cure-all which the businessman can buy for a fixed price and put into operation with no further thought. The principal limitations of the technique today lie in three areas:

1. *Cost or revenue proportional to volume*—Problem-solving procedures have been well developed only for problems where the cost incurred or revenue produced by every possible activity is strictly proportional to the volume of that activity; these are the procedures that belong under the somewhat misleading title of *linear* programing. This limitation, however, is not so serious as it seems. Problems involving nonproportional costs or revenues can often be handled by linear programing through the use of special devices or by suitable approximations, and research is progressing on the development of procedures which will handle some of these problems directly.

2. *Arithmetic capacity*—Even when the procedure for solving a problem is perfectly well known, the solution may involve such a sheer quantity of arithmetic that it is beyond the capacity even of electronic computing machinery. However, the problem can sometimes be set up more simply so that solution is practical. For instance, careful analysis may show that the really essential variables are relatively few in number, or that the problem may be split into parts of manageable size.

3. *Scheduling problems*—A third limitation is often the most serious, particularly in the assignment of machine tools. So far very little has been accomplished toward the solution of *scheduling* problems, where certain operations must be performed before or after other operations. Mathematical programing can indicate, within the limits of available tool capacity, which operations should be performed on which tools, but the arrangement of these operations in the proper sequence must usually be handled as a separate problem. Again, however, research is attempting to find procedures which will reduce even this problem to a straight-forward routine, and some progress in this direction has already been made.

APPLICATION

In Part II of this article we describe a series of cases which should suggest to the reader the sort of problems where mathematical programing can be of use in his own business. Included are both actual cases and hypothetical examples. The hypothetical examples are purposely made so simple that they could be solved *without* the use of these procedures; in this way the reader can better see the essential nature of the analysis which programing will accomplish in more complex problems.

Top executives may want to turn a detailed reading of this section over to specialists, but they will find the major points as set forth below of practical interest. Very briefly, the discussion of case examples will show that mathe-

matical programing can be used to decide:

1. *Where to ship*—Here the problem is to find the shipping program that will give the lowest freight costs. It has been demonstrated by the H. J. Heinz Company that linear programing can save thousands of dollars on a single scheduling problem alone. By virtue of its greater ease and accuracy, linear programing has also enabled the company to schedule on a monthly rather than quarterly basis, thus taking advantage of new information as soon as it becomes available.

2. *Where to ship and where to produce*—A complete program to determine the most economical program of production or procurement *and* freight costs can be developed so quickly and inexpensively that every possible alternative can be taken into account without throwing a heavy burden on senior personnel.

3. *Where to ship, where to produce, and where to sell*—Here the problem is further complicated. Such factors as a management policy regarding minimum supplies for dealers and a varying price schedule should and can be taken into account.

4. *What the most profitable combination of price and volume is*—At present mathematical programing can provide the answers only under certain conditions, but progress is being made in broadening its applicability.

5. *What products to make*—Problems that can be solved range from the most economical use of scarce raw materials to the most profitable mix in gasoline blending. If automatic computers are necessary because of the sheer bulk of arithmetic, the small or medium-size firm can turn to a central service bureau; the company does not have to be so large that it can afford its own computers.

6. *What products to make and what processes to use*—This problem arises when machine capacity is limited. Here mathematical programing may produce surprising results. For example, a certain amount of idle time on one machine may be necessary for the greatest production. Without mathematical programing, there is a real danger that personnel will use every machine all the time to satisfy management pressure, and thus defeat the company's real objective.

7. *How to get lowest cost production*—Here the problem is to determine the most economical production when the company can produce all it can sell. In these days of growing cost-consciousness, mathematical programing may become one of management's really valuable cost-reduction tools.

The businessman who recognizes or suspects that he has a problem which can be solved by mathematical programing will usually have to consult with specialists to learn how to use the technique. But an even greater responsibility will remain with the businessman himself. Like the introduction of a variable overhead budget, each application of mathematical programing will require careful study of the particular circumstances and problems of the company involved; and, once installed, the technique will pay off only in proportion to the understanding with which management makes use of it.

PART II. EXAMPLES OF OPERATION

The case examples to be presented here illustrate some of the uses of mathematical programing. Although limited in number, the examples are so arranged that the reader who follows

them through in order should gain an understanding of the situations in which mathematical programing can and cannot be helpful and of how to set up any problem for accurate solution. The exhibits accompanying the text set forth the mathematical solution of the problems posed in the cases, while the appendix gives specific directions on how to work through a procedure for handling some of the problems that may arise in the reader's own business.

WHERE TO SHIP

As our first example of the uses of mathematical programing, let us look at a case where the technique is currently in use as a routine operating procedure in an actual company:

The H. J. Heinz Company manufactures ketchup in half a dozen plants scattered across the United States from New Jersey to California and distributes this ketchup from about 70 warehouses located in all parts of the country.

In 1953 the company was in the fortunate position of being able to sell all it could produce, and supplies were allocated to warehouses in a total amount exactly equal to the total capacity of the plants. Management wished to supply these requirements at the lowest possible cost of freight; speed of shipment was not important. However, capacity in the West exceeded requirements in that part of the country, while the reverse was true in the East; for this reason a considerable tonnage had to be shipped from western plants to the East. In other words, the cost of freight could not be minimized by simply supplying each warehouse from the nearest plant.

Simplest Problem This problem can immediately be recognized as a problem of *programing* because its essence is the minimization of cost subject to a fixed set of plant capacities and warehouse requirements. It can be handled by *linear* programing because the freight bill for shipments between any two points will be proportional to the quantity shipped. (The quantities involved are large enough so that virtually everything will move at carload rates under *any* shipping program which might be chosen.)

This is, in fact, the simplest possible kind of problem that can be solved by this method. Certain complexities which make solution by trial and error considerably more difficult than usual—in particular, the existence of water-competitive rates, which make it practical to send California ketchup all the way to the East Coast—add no real difficulty to the solution by linear programing. Given the list of plant capacities and warehouse requirements, plus a table of freight rates from every plant to every warehouse, one man with no equipment other than pencil and paper solved this problem for the first time in about 12 hours. After H. J. Heinz had adopted the method for regular use and clerks had been trained to become thoroughly familiar with the routine for this particular problem, the time required to develop a shipping program was considerably reduced.

The actual data of this problem have not been released by the company, but a fair representation of its magnitude is given by the similar but hypothetical examples of Tables 1 and 2, which show the data and solution of a problem of supplying 20 warehouses from 12 plants.

Table 1 shows the basic data: the body of the table gives the freight rates, while the daily capacities of the plants and daily requirements of the warehouses are in the margins. For example, Factory III, with a capacity of 3,000 cwt. per day, can supply Warehouse G, with requirements of 940 cwt. per day, at a freight cost of 7 cents per cwt.

Any reader who wishes to try his hand will quickly find that without a systematic procedure a great deal of work would be required to find a shipping program which would come reasonably close to satisfying these requirements and capacities at the lowest possible cost. But with the use of linear programing the problem is even easier than the Heinz problem.

Table 2 gives the lowest-cost distribution program. For example, Warehouse K is to get 700 cwt. per day from Factory I and 3,000 cwt. per day from Factory III. On the other hand, Factory III ships nothing to Warehouse A, although Table 1 shows that Factory III could ship at less expense to this warehouse than to any other. . . .

TABLE 1

TABLE OF RATES, REQUIREMENTS, AND CAPACITIES

Factory	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*	*Daily requirements (cwt.)*
	Freight rates (cents per cwt.)												
Warehouse A	16	16	6	13	24	13	6	31	37	34	37	40	1,820
B	20	18	8	10	22	11	8	29	33	25	35	38	1,530
C	30	23	8	9	14	7	9	22	29	20	38	35	2,360
D	10	15	10	8	10	15	13	19	19	15	28	34	100
E	31	23	16	10	10	16	20	14	17	17	25	28	280
F	24	14	19	13	13	14	18	9	14	13	29	25	730
G	27	23	7	11	23	8	16	6	10	11	16	28	940
H	34	25	15	4	27	15	11	9	16	17	13	16	1,130
J	38	29	17	11	16	27	17	19	8	18	19	11	4,150
K	42	43	21	22	16	10	21	18	24	16	17	15	3,700
L	44	49	25	23	18	6	13	19	15	12	10	13	2,560
M	49	40	29	21	10	15	14	21	12	29	14	20	1,710
N	56	58	36	37	6	25	8	19	9	21	15	26	580
P	59	57	44	33	5	21	6	10	8	33	15	18	30
Q	68	54	40	38	8	24	7	19	10	23	23	23	2,840
R	66	71	47	43	16	33	12	26	19	20	25	31	1,510
S	72	58	50	51	20	42	22	16	15	13	20	21	970
T	74	54	57	55	26	53	26	19	14	7	16	6	5,110
U	71	75	57	60	30	44	30	30	41	8	23	37	3,540
Y	73	72	63	56	37	49	40	31	31	10	8	25	4,410
Daily capacity (cwt.)	10,000	9,000	3,000	2,700	500	1,200	700	300	500	1,200	2,000	8,900	40,000

Advantages Gained One of the most important advantages gained by the H. J. Heinz Company from the introduction of linear programing was relief of the senior members of the distribution department from the burden of preparing shipping programs. Previously the quarterly preparation of the program took a substantial amount of their time; now they pay only as much attention to this problem as they believe necessary to keep the feel of the situation, while the detailed development of the program has been handed over to clerks. Freed from the burden of working out what is after all only glorified arithmetic, they have this much more time to devote to matters which really require their experience and judgment.

TABLE 2

LOWEST-COST DISTRIBUTION PROGRAM (DAILY SHIPMENTS FROM FACTORY TO WAREHOUSE IN CWT.)

Factory	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*	*Row Total*	*Row value*
Warehouse														
A	1,820												1,820	16
B	1,530												1,530	20
C		2,360											2,360	28
D	100												100	10
E		280											280	28
F		730											730	19
G	940												940	27
H				1,130									1,130	28
J		4,150											4,150	34
K	700		3,000										3,700	42
L	1,360					1,200							2,560	44
M		140		1,570									1,710	45
N	580												580	56
P								30					30	51
Q		1,340			500				500			500	2,840	59
R	810						700						1,510	66
S								90				880	970	57
T												5,110	5,110	42
U	2,160							180		1,200			3,540	71
Y											2,000	2,410	4,410	61
Total	10,000	9,000	3,000	2,700	500	1,200	700	300	500	1,200	2,000	8,900	40,000	
Column value	0	−5	−21	−24	−51	−38	−54	−41	−49	−63	−53	−36		

An equally important gain, in the opinion of these officials themselves, is the peace of mind which results from being sure that the program *is* the lowest-cost program possible.

The direct dollars-and-cents saving in the company's freight bill was large

enough by itself to make the use of this technique very much worth while. The first shipping program produced by linear programing gave a projected semiannual freight cost several thousand dollars less than did a program prepared by the company's previous methods, and this comparison is far from giving a full measure of the actual freight savings to be anticipated.

Shipping schedules rest on estimates which are continuously subject to revision. The capacity figures in part represent actual stocks on hand at the plants, but in part they are based on estimates of future tomato crops; and the figures for requirements depend almost wholly on estimates of future sales. The fact that schedules are now quickly and accurately prepared by clerks has enabled the company to reschedule monthly rather than quarterly, thus making much better use of new information on crops and sales as it becomes available.

Furthermore, the risk of backhauling is very much reduced under the new system. It had always been company practice early in the season to hold "reserves" in regions of surplus production, in order to avoid the danger of shipping so much out of these regions that it became necessary to ship back into them when production and sales estimates were revised. In fact, these reserves were largely accidental leftovers: when it became really difficult to assign the last part of a factory's production, this remainder was called the reserve. Now the company can look at past history and decide in advance what reserve should be held at each factory and can set up its program to suit this estimate exactly. Since the schedule is revised each month, these reserves can be altered in the light of current information until they are finally reduced to nothing just before the new pack starts at the factory in question.

Similar Problems Many important problems of this same character unquestionably are prevalent in business. One such case, for instance, would be that of a newsprint producer who supplies about 200 customers all over the United States from 6 factories scattered over the width of Canada.[1]

Similar problems arise where the cost of transportation is measured in time rather than in money. In fact, the first efforts to solve problems of this sort systematically were made during World War II in order to minimize the time spent by ships in ballast. Specified cargo had to be moved from specified origins to specified destinations; there was usually no return cargo, and the problem was to decide to which port the ship should be sent in ballast to pick up its next cargo. An obviously similar problem is the routing of empty freight cars,[2] and a trucker operating on a nationwide scale might face the same problem with empty trucks.

WHERE TO PRODUCE

When ketchup shipments were programed for the H. J. Heinz Company, factory capacities and warehouse requirements were fixed before the ship-

[1] R. Dorfman, "Mathematical, or 'Linear' Programming," *American Economic Review*, Dec. 1953, p. 797.

[2] Cf. *Railway Age*, April 20, 1953, pp. 73–74.

ping program was worked out, and the only cost which could be reduced by programing was the cost of freight. Since management had decided in advance how much to produce at each plant, *all* production costs were "fixed" *so far as the programing problem was concerned.*

The same company faces a different problem in connection with another product, which is also produced in a number of plants and shipped to a number of warehouses. In this case, the capacity of the plants *exceeds* the requirements of the warehouses. The cost of production varies from one plant to another, and the problem is thus one of satisfying the requirements at the least *total* cost. It is as important to reduce the cost of *production* (by producing in the right place) as it is to reduce the cost of *freight* (by supplying from the right place). In other words, management must now decide two questions instead of one: (a) How much is each factory to produce? (b) Which warehouses should be supplied by which factories?

It is tempting to try to solve these two problems one at a time and thus simplify the job, but in general it will *not* be possible to get the lowest total cost by first deciding where to produce and then deciding where to ship. It is obviously better to produce in a high-cost plant if the additional cost can be more than recovered through savings in freight.

Method of Attack This double problem can be handled by linear programming if we may assume (as businessmen usually do) that the cost of production at any one plant is the sum of a "fixed" cost independent of volume and a "variable" cost proportional to volume in total but fixed per unit, and if these costs are known. The variable cost is handled directly by the linear programing procedure, while the fixed part is handled by a method which will be explained later.

Actually, the problem can be much more complicated and still lend itself to solution by linear programing. For example, we can bring in the possibility of using overtime, or of buying raw materials at one price up to a certain quantity and at another price beyond that quantity.

Table 3 shows the cost information needed to solve a hypothetical example of this sort. It is assumed that there are only four plants and four warehouses, but any number could be brought into the problem.

In our first approximation (which we shall modify later) we shall assume that no plant will be closed down entirely and, therefore, that "fixed costs" are really fixed and can be left out of the picture. Like Table 1, Table 3 shows the freight rates from each plant to each warehouse, the available daily capacity at each plant, and the daily requirements of each warehouse; it also shows the "variable" (fixed-per-unit) cost of normal production at each plant and the additional per-unit cost of overtime production. The total capacity is greater than the total requirements even if the factories work only normal time.

On the basis of these data, the lowest-cost solution is given by Part A of Table 4. It is scarcely surprising that this

TABLE 3

COST INFORMATION FOR DOUBLE PROBLEM

A — Warehouse Requirements (tons per day)					
Warehouse	*A*	*B*	*C*	*D*	*Total*
Requirements	90	140	75	100	405
B — Factory Capacities (tons per day)					
Factory	*I*	*II*	*III*	*IV*	*Total*
Normal capacity	70	130	180	110	490
Additional capacity on overtime	25	40	60	30	155
C — Variable Costs (per ton)					
Factory	*I*	*II*	*III*	*IV*	
Normal production cost	$30	$36	$24	$30	
Overtime premium	15	18	12	15	
Freight rates to:					
Warehouse A	$14	$ 9	$21	$18	
B	20	14	27	24	
C	18	12	29	20	
D	19	15	27	23	

solution calls for no use of overtime. So long as fixed costs are taken as really fixed, it turns out that it is best to use the entire normal capacity of Factories I, II, and III, and to use 25 tons of Factory IV's normal capacity of 110 tons per day. The remaining 85 tons of normal capacity at IV are left unused. The total variable cost under this schedule (freight cost plus variable production cost) will be $19,720 per day.

Final Determination Presented with this result, management would certainly ask whether it is sensible to keep all four factories open when one of them is being left about 80% idle. Even without incurring overtime, Factory I, the smallest plant, could be closed and the load redistributed among the other plants. If this is done, the lowest-cost distribution of the requirements among Factories II, III, and IV is that given by Part B of Table 4. Under this program the total variable cost would be $19,950 per day, or $230 per day more than under the program of Table 4, A, which depended on the use of all four plants. If more than $230 per day of fixed costs can be saved by closing down Factory I completely, it will pay to do so; otherwise it will not.

It might be still better, however, to close down some plant other than Fac-

TABLE 4

LOWEST-COST DISTRIBUTION PROGRAM (DAILY SHIPMENTS IN TONS FROM FACTORY TO WAREHOUSE)

A — With All Four Factories Open

Factory	*I*	*II*	*III*	*IV*	*Total*
Warehouse A			90		90
B		80	60		140
C		50		25	75
D	70		30		100
Idle normal capacity				85	85
Total	70	130	180	110	490

B — With Factory I Closed

Factory	*II*	*III*	*IV*	*Total*
Warehouse A		90		90
B	130	10		140
C			75	75
D		80	20	100
Idle normal capacity			15	15
Total	130	180	110	420

C — With Factory IV Closed

Factory	*I*	*II*	*III*	*Total*
Warehouse A			90	90
B		55	85	140
C		75		75
D	70		30	100
Total	70	130	205	405

tory I even at the cost of a certain amount of overtime. In particular, a very little overtime production (25 tons per day) would make it possible to close Factory IV. A person asked to look into this possibility might reason as follows: Under the shipping schedule of Table 4, A, the only use of Factory IV's capacity is to supply 25 tons per day to Warehouse C. Looking at Table 3 for a replacement for this supply, he would get the following information on costs per ton:

Factory	*Normal cost of production*	*Overtime premium*	*Freight to Warehouse C*	*Total*
I	$30	$15	$18	$63
II	36	18	12	66
III	24	12	29	65

Apparently the cheapest way of using overtime, if it is to be used at all, would be to produce the needed 25 tons per day at Factory I and ship them to Warehouse C at a total varia-

ble cost of $63 per ton. Under the program of Table 4, A, with all plants in use, Warehouse C was supplied from Factory IV at a total variable cost of $30 for production plus $20 for freight, or a total of $50 per ton. The change would thus seem to add a total of $325 per day (25 tons times $13 per ton which is the difference between $63 and $50 per ton).

But, in fact, closing Factory IV need not add this much to the cost of the program. If we take Factory IV out of the picture and then program to find the best possible distribution of the output of the remaining plants, we discover that the program of Part C of Table 4 satisfies all requirements at a total variable cost of $19,995 per day, or only $275 per day more than with all plants in use. The overtime is performed by Factory III, which does not supply Warehouse C at all.

Difficulties Avoided This last result deserves the reader's attention. *Once a change was made in a single part of the program, the best adjustment was a general readjustment of the entire program.* But such a general readjustment is impractical unless complete programs can be developed quickly and at a reasonable cost. It is rarely clear *in advance* whether the work will prove profitable, and management does not want to throw a heavy burden of recalculation on senior personnel every time a minor change is made. Mathematical programing avoids these difficulties. Even minor changes in the data can be made freely despite the fact that complete recalculations of the program are required, because the work can be done quickly and accurately by clerks or machines.

We can proceed to compute the lowest possible cost of supplying the requirements with Factory II or Factory III closed down completely. We can then summarize the results for all alternatives like this:

	Total freight plus variable production cost
All four factories in use	$19,720
Factory I closed, no overtime	19,950
Factory II closed, overtime at Factory III	20,515
Factory III closed, overtime at Factories I, II, and IV	21,445
Factory IV closed, overtime at Factory III	19,995

Management now has the information on variable costs which it needs in order to choose rationally among three alternatives: (1) operating all four plants with a large amount of idle normal capacity; (2) shutting down Factory I and still having a little idle normal capacity; (3) shutting down Factory II, III, or IV and incurring overtime. Its choice will depend in part on the extent to which fixed costs can be eliminated when a particular plant is completely closed; it may depend even more on company policies regarding community relations or some other nonfinancial consideration. Mathematical programing cannot replace judgment, but it can supply some of the factual information which management needs in order to make judgments.

Related Problems Problems of this general type are met in purchasing as well as in producing and selling. A company which buys a standard raw

material at many different geographical locations and ships it to a number of scattered plants for processing will wish to minimize the total cost of purchase plus freight; here the solution can be obtained in exactly the same way as just discussed. The Department of Defense is reported to have made substantial savings by using linear programing to decide where to buy and where to send certain standard articles which it obtains from a large number of suppliers for direct shipment to military installations.

WHERE TO SELL

In our first case, we considered a situation where management had fixed the sales at each warehouse and the production at each plant before using programing to work out the best way of shipping from plant to warehouse. In the second example, management had fixed the sales at each warehouse in advance, but had left the decision on where and how much to produce to be made as a part of the program. Let us now consider a case in which sales are not fixed in advance, and management wants to determine where to sell, as well as where to produce and where to ship, in order to give the greatest possible profits.

Such a problem often arises when sales would exceed a company's capacity to produce unless demand were retarded by higher prices, yet management does not wish to raise prices because of the long-run competitive situation. Under these circumstances some system of allocating the product to branch warehouses in the different market areas (or to individual customers) will be necessary. One way of doing this is simply to sell wherever the greatest short-run profits can be made. Often, however, management will not want to take an exclusively short-run view and will want to provide each warehouse or customer with at least a certain minimum supply, with only the remainder over and above these minimum allocations being disposed of with a view to maximum short-run profits.

One additional complication will often be present in real problems of this sort. The selling price of the product may not be uniform nationally, but may vary from place to place or from customer to customer. In addition, there may well be present the complication we dealt with in the last example: it may be desirable to have some plants working overtime while others are working at only a part of their normal capacity or are even closed down entirely.

Thus a production and distribution program must be prepared which answers all the following questions in such a way as to give the greatest possible profits, subject to the requirement of supplying certain warehouses with at least a specified allocation of product:

(1) How much shall be produced at each plant?

(2) How much, if any, above the predetermined minimum shall be delivered to each warehouse?

(3) The above questions being answered, which plants shall supply which warehouses?

As in the previous example, all three questions must be answered simultaneously; it is not possible to work them

out one by one. The problem can still be handled by linear programing, however, despite the additional complications which have entered the picture; in fact, it is no harder to solve than the previous problem. The only difference is that we now look directly at the profit resulting from supplying a particular warehouse from a particular plant, rather than looking at the costs involved. We shall not even work out an example, since the solution would appear in the same form as Table 4 of the previous case, while the required data would look the same as Table 3 with the addition of the selling price at each warehouse.

PRICE, VOLUME, AND PROFIT

In all the previous examples it was assumed that management had set selling prices before the production and distribution program was worked out. The quantity to be produced and shipped followed from the predetermined prices. This is certainly a common situation, but it is also very common for management to want to consider the effect of prices on volume *before* prices are set. This means, of course, that sales volume must be forecast at each of a variety of possible prices, and we assume that such forecasts have been made separately for each of the branch warehouses of our previous examples.

Under these conditions the problem can no longer be handled *directly* by linear programing, since the margin, or difference between the selling price at a particular warehouse and the variable cost of producing at a particular plant and shipping to that warehouse, is no longer in a constant ratio to the quantity produced and sold. As quantities go up, prices go down, and the ratio of total margin to quantity sold declines. Even so, we can still use linear programing to solve the problem quickly, accurately, and cheaply *if* there is to be a single national selling price. We can compute the best program for each proposed price, determine the total profits for each program, and select the most profitable alternative.

However, linear programing becomes virtually impossible if prices can vary from place to place and management wishes to set each local price in such a way as to obtain the greatest total profits. Even if there are only ten distribution points for which price-quantity forecasts have to be considered, and even if each branch manager submits forecasts for only five different prices, we would have to compute nearly 10 million different programs and then select the most profitable one.

In practical cases it will often prove possible with a reasonable amount of calculation to find a program which is probably the best program or very close to it, but in general the solution of this problem of mathematical programing, like many others, depends on further research to develop methods for attacking nonlinear problems directly. As mentioned, progress in this direction is already being made.

WHAT AND HOW TO PRODUCE

All the cases discussed so far have involved problems of *where* (as well as how much) to buy, sell, produce, and ship. Mathematical programing can be

of equal use in deciding *what* and *how* to produce in order to maximize profits or minimize costs in the face of shortages of raw materials, machine tools, or other productive resources. Some problems of this kind may be solved by clerks using procedures such as those previously discussed; others, however, may require new procedures and automatic computing equipment.

A representative problem in the first category is the following one, which involves the selective use of scarce raw materials:

A manufacturer produces four products, A, B, C, and D, from a single raw material which can be bought in three different grades, I, II, and III. The cost of processing and the quantity of material required for one ton of end product vary according to the product and the grade of material used, as shown in Table 5.

If unlimited supplies of each grade of material were available at a fixed market

TABLE 5

COSTS, AVAILABILITIES, AND PRICES

A — Yields and Processing Costs

Grade	*I*	*II*	*III*
Product	*Tons of material per ton of product*		
A	1.20	1.80	2.00
B	1.50	2.25	2.50
C	1.50	2.25	2.50
D	1.80	2.70	3.00
	Processing cost per ton of product		
A	$18	$30	$ 42
B	30	60	69
C	57	63	66
D	54	81	126

B — Material Cost and Availability

Grade	*I*	*II*	*III*
Normal price per ton	$48	$24	$18
Quantity available at normal price (tons)	100	150	250
Premium price per ton	$72	$36	$24
Quantity available at premium price (tons)	100	150	400

C — Product Prices and Sales Potentials

Product	*A*	*B*	*C*	*D*
Price per ton	$96	$150	$135	$171
Potential sales (tons)	200	100	160	50

price, each product would be made from the grade for which the total purchasing-plus-processing cost was the smallest; but the amount of each grade obtainable at the "normal" price is limited as shown in the exhibit. Additional quantities of any grade can be obtained, but only at the premium shown.

The products are sold f.o.b. the manufacturer's single plant; the selling prices have already been set and are shown in the exhibit, together with the sales department's forecasts of the amount of each product which can be sold at these prices.

TABLE 6

MOST PROFITABLE PRODUCTION PROGRAM

	Tons of product		*Tons of material used*		
Product	*Sales potential*	*Production*	*Grade I*	*Grade II*	*Grade III*
A	200	200		210	167
B	100	100	100		83
C	160	160			400
D	50	0			
Total material usage			100	210	650
Bought at normal price			100	150	250
Bought at premium price			0	60	400

The problem, then, is to determine *what* products to make and how much of each, and *how* to make them—in other words, which grade of material to use for which products. The solution is shown in Table 6.

Use of Computers It will be remembered that, in discussing the use of mathematical programing by the H. J. Heinz Company, we emphasized the fact that shipping programs are produced by a clerk with nothing but paper and pencil in a very reasonable amount of time. This is true even though the existence of about 6 plants and 70 warehouses makes it necessary to choose 75 routes for actual use from the 420 possible routes which might be used. This ease of solution, even in cases where a very large number of variables is involved, applies to the selective use of raw materials just discussed as well as to the other problems taken up in earlier sections. They are all problems which can be solved by what is known as the "transportation-problem procedure."

By contrast, other problems usually require the use of high-speed computing machinery. They are problems requiring the use of what might be called the "general procedure." While the mathematics involved here is at the level of grade-school arithmetic, the sheer *bulk* of arithmetic required is very much greater than under the transportation procedure. This means that, unless a skilled mathematician finds some way of simplifying a particular problem, it will be impossible for clerks to obtain

a solution by hand in a reasonable amount of time when the number of variables is such as will be encountered in most practical situations.

Whether a given problem can be solved by the transportation-problem procedure or will require the use of the general procedure does not depend on whether the problem actually involves transportation or not, but rather on the *form* of the data. The raw material problem discussed just above, for example, could be solved as a transportation problem because *any* product would require 50% more material if Grade II was used instead of Grade I, or 67% more if Grade III was used instead of Grade I. But if the inferiority of yield of the lower grades had varied depending on the particular end product, it would have been necessary to use the general procedure.

The fact that the general procedure usually requires an automatic computer by no means implies that this procedure can be profitably applied only by very large firms with computers of their own. Fortunately, all problems which call for the use of this procedure are *mathematically* the same, even though the physical and economic *meaning* of each problem may be completely different. And since they are mathematically the same, a machine at a central service bureau can be coded once and for all to carry out the general procedure for any problem up to a certain size. The machine can then be used to solve the varying problems of many different companies promptly and inexpensively. Such a service can already be purchased from at least one source by the hour, and the time required to solve a problem is usually surprisingly short.

Most Profitable Blend Now let us turn to a case requiring the use of the general procedure:

> Gasoline sold as an automobile or aviation fuel is ordinarily not the product of a single refining process but a blend of various refinery products with a certain amount of tetraethyl lead added. To a certain extent each of the various constituents requires peculiar refining facilities. Consequently, the management of a refinery may well be faced with the following problem: given a limited daily supply of each of various constituents, into what end-product fuels should they be blended to bring in the maximum profits? The problem is made additionally complicated by the fact that there is no single "recipe" for any particular end product. In general, the end product may be blended in any of a large number of different ways; provided only that certain performance specifications are met.

This is clearly a problem of programing, both because the use of a given constituent in one end product means that less is available for use in another, and also because the use of one constituent to produce a given kind of performance in a particular end product means that less of other constituents is needed to produce that performance in the end product. But is the problem linear? We must look a little more closely at the relation between the characteristics of the constituents and the characteristics of the resulting blends:

> The two most important measures of the performance characteristics of a gasoline fuel are its performance number (PN),

which is a development of the octane number and describes antiknock properties, and its vapor pressure (RVP), which indicates the volatility of the fuel. In the case of most high-grade aviation gasolines there are actually two PN's specified: the 1-c PN, which applies to lean mixture, and the 3-c PN, which applies to rich mixture. Each of the various constituents has its own RVP and PN.

The PN and RVP required in the end product are produced by proper blending of the constituents and by the addition of tetraethyl lead (TEL) to improve the PN. The amount of TEL which can be used in any fuel is limited for various reasons; and since TEL is often the cheapest way of obtaining the desired PN (particularly in the case of aviation fuels), it is a common practice to use the maximum permitted amount of this chemical.

It appears from the above that the problem will be linear provided that the RVP and PN of any end product are simply weighted averages of the RVP's and PN's of the various constituents (each PN being calculated for the predetermined amount of TEL to be used in the end product). While not perhaps strictly true as regards PN, this proposition is close enough to the truth to serve as the basis for ordinary blending calculations. Therefore the problem can be handled in a straightforward manner by linear programing.

A. Charnes, W. W. Cooper, and B. Mellon have applied linear programing to the choice of the most profitable mix in an actual refinery; and although they were forced to simplify the problem somewhat in order to do the computation with nothing but a desk calculator, the results of their calculations were of considerable interest to the company's management.[3] (With modern computing equipment, of course, much more data could be handled in much less time, and various large oil companies are currently trying out the use of such equipment for this purpose.)

The figures which Charnes, Cooper, and Mellon present to show the nature of the calculations, and which we use below, are of course largely disguised:

The refinery in question is considered as having available fixed daily supplies of one grade of each of four blending constituents: alkylate, catalytic-cracked gasoline, straight-run gasoline, and isopentane. The quantities available and the performance specifications are shown in Table 7. These constituents can be blended into any of three different aviation gasolines, A, B, or C, the specifications and selling prices of which are also shown in Table 7.

Any supplies not used in one of these three aviation gasolines will be used in premium automobile fuel, the selling price of which likewise appears in the exhibit. Performance specifications for automobile fuel are not shown since this product will be composed primarily of constituents not included in this study; these constituents will be added in the proper proportions to give the desired performance specifications.

Management has decided to use the entire available supply of the constituents in one way or another. Their costs can therefore be neglected in selecting the blending program since they will be the same whatever program is chosen. The costs of blending itself are also about the same whatever end product is produced and can, there-

[3] A. Charnes, W. W. Cooper, and B. Mellon, "Blending Aviation Gasolines—A Study in Programming Interdependent Activities in an Integrated Oil Company," *Econometrica*, April 1952, p. 135.

TABLE 7

QUANTITIES AVAILABLE AND PERFORMANCE SPECIFICATIONS

A — Product Specifications

Product	*Maximum RVP*	*Minimum 1-c PN*	*Minimum 3-c PN*	*Maximum TEL cc. per gal. of product*	*Price per bbl. of product*	*Cost of TEL per bbl. of product*
Avgas A	7.0	80.0	—	0.5	$4.960	$0.051770
Avgas B	7.0	91.0	96.0	4.0	5.846	0.409416
Avgas C	7.0	100.0	130.0	4.0	6.451	0.409416
Automobile	—	—	—	3.0	4.830	0.281862

B — Constituent Specifications

Constituent	*Supply bbl. per day*	*RVP*	*1-c PN* 0.5 cc. TEL	*1-c PN* 4.0 cc. TEL	*3-c PN* 4.0 cc. TEL
Alkylate	3,800	5.0	94.0	107.5	148.0
Catalytic	2,652	8.0	83.0	93.0	106.0
Straight-run	4,081	4.0	74.0	87.0	80.0
Isopentane	1,300	20.5	95.0	108.0	140.0

fore, be neglected in solving this problem, too. The only variable cost factor is the TEL (since some end products use more of this than others), and its cost per barrel of product is shown in Table 7.

The solution of the problem is given in Table 8. In the actual case, however, precise determination of the most profitable blending program was not the result which was of most interest to the management concerned. After all, the company's experienced schedulers *could*, given sufficient time, arrive at programs as profitable or nearly as profitable as those derived by mathematical programing—although the tests which seemed to show this were perhaps unduly favorable to the traditional methods because the schedulers were given the results of the programing calculations in advance, and thus knew what they had to try to attain.

The indirect results were what really impressed management. For one thing, just as in the case of the Heinz Company, it was clear that the time and effort of experienced personnel would be saved if the job were routinized by the use of mathematical programing. This, in turn, now made it practical to compute programs for a variety of requirements and assumptions not previously covered. To illustrate:

The most profitable product mix as shown in Table 8 contains no Avgas A. However, company policy called for the production of 500 bbl. per day of this product for goodwill reasons. When the problem was recomputed taking this factor into account, it was found that the most

profitable mix containing the required 500 bbl. of Avgas A yielded profits about $80,000 per year less than those resulting from the program of Table 8.

This loss was considerably higher than management had believed. Presumably the cost could have been computed with adequate accuracy by the company's schedulers, but *when such calculations are expensive in terms of the time of senior personnel, they simply do not get made.*

inapplicable when the problem involves the blending of automotive rather than high-grade aviation fuels. In such a case it is not at all clear in advance that it will be economical to use the maximum permitted amount of TEL, and PN is definitely not proportional to the amount of TEL in the fuel.

The procedures which have been developed to cope with situations like

TABLE 8

MOST PROFITABLE PRODUCT MIX

Product	*Total amount produced*	*Composed of these constituents:*			
		Alkylate	*Catalytic*	*Straight-run*	*Isopentane*
Avgas A	0	0	0	0	0
Avgas B	5,513	0	2,625	2,555	333
Avgas C	6,207	3,800	27	1,526	854
Automobile	113	0	0	0	113
Total	11,833	3,800	2,652	4,081	1,300

"Concave" Programing The field of gasoline refining is perhaps the one in which the most extensive work has been done in trying out actual applications of mathematical programing to practical operations. One interesting type of *non*linear programing has been tried on actual data in this field. The method has been called "concave" programing.

In our gasoline case, the problem could be solved by linear programing because it was assumed that the RVP and PN of any product would be a simple weighted average of the RVP's and PN's of the constituents, the PN's being calculated for a predetermined amount of TEL in the product. We have already suggested that under some conditions this assumption is not strictly true. Linear programing is particularly

this have at least approximately solved the problem in a number of actual cases.[4] The results show the most profitable amount of TEL to use in various end products as well as the most profitable way to blend the refinery stocks.

WHAT PROCESSES TO USE

Some of the most perplexing problems of limited resources which management commonly faces do not concern materials but the productive capacity of the plant. A good example is the problem of choosing what products to make and what processes to use for manufacturing them when a shortage

[4] See A. S. Manne, *Concave Programming for Gasoline Blends*, Report P-383 of The Rand Corporation, Santa Monica, 1953.

of machine capacity restricts production. The problem may arise because of a shortage of only a few types of machine in a shop which is otherwise adequately equipped. The SKF Company, for example, has reported savings of $100,000 a year through the use of scheduling techniques developed from linear programing.[5]

Rather than describing the SKF application, however, let us take a hypothetical example which will give an opportunity to show one of the ways in which setup costs can be handled by mathematical programing. Setup costs cannot be handled directly by linear programing because they are not proportional to volume of production. However, they can be handled indirectly by the same means used to deal with the fixed costs that can be avoided by closing down a plant completely (see the case described under the heading Where to Produce). Here is an illustrative situation:

A machine shop has adequate machine-tool capacity except for three types of machine, I, II, and III. These machines are used (in conjunction with others) to make three products, A, B, and C. Each product can be made in a variety of ways. It is possible, for example, to reduce the amount of time required for grinding by closer machining, but this requires more machining time. To be specific, let us suppose that for each product there are three alternate operation sheets, which we shall call processes 1, 2, and 3.

If sufficient time were available on all machines, the most economical process would be chosen for each product individually, and the company would then make all it could sell of that product. But because of the shortage of capacity the process to be used for any one product must be chosen with regard to its effect on machine availability for the other two products, and the quantity to be produced must be calculated for all products together in such a way as to obtain the greatest profit from the total production of all products.

The demands of each process for each product on the three critical types of machine are shown in Table 9; these are per-unit times (standards duly adjusted for efficiency). For example, if Product B is produced by Process 3, each unit will require 0.2 hour on a machine of Type II and 1.0 hour on a machine of Type III, but no time on Type I. The weekly available machine hours are also shown in the table, after deduction of estimated allowances for repair and maintenance, but with no deduction for setup.

Table 9 also shows the number of units of each product which must be produced each week to fill orders already accepted, together with the "margin" which will be realized on any additional units that can be produced. This margin is the selling price less all out-of-pocket costs of production *except* the costs of operating the machines being programed. Since these machines are the "bottlenecks," they will be used full time or virtually full time in any case, and, therefore, the costs of operating them will be virtually the same regardless of the program chosen.

Solution of Problem To solve the problem, we start by neglecting the setup times for the machines (shown in Table 9) just as we first neglected fixed costs in deciding where to produce ketchup. We simply deduct a roughly estimated flat six hours from each of the

[5] *Factory Management and Maintenance,* January 1954, pp. 136–137. The technique there described is very close to the "profit-preference procedure" mentioned in the Appendix.

TABLE 9

MACHINE-SHOP REQUIREMENTS

A — Per-Unit Machine Times

Machine type		*I*	*II*	*III*
Product	*Process*	*Machine hours per unit*		
A	1	0.2	0.2	0.2
A	2	0.4	—	0.3
A	3	0.6	0.1	0.1
B	1	0.2	0.3	0.4
B	2	0.1	0.1	0.8
B	3	—	0.2	1.0
C	1	0.2	0.1	0.7
C	2	0.1	0.6	0.4
C	3	—	0.8	0.2

B — Total Machine Hours Available per Week

Machine type	*I*	*II*	*III*
Hours	118	230	306

C — Product Requirements and "Margins"

Product	*A*	*B*	*C*
Minimum units required per week	100	200	300
Margin per unit on additional production	$10	$20	$30

D — Machine Setup Times

Machine type		*I*	*II*	*III*
Product	*Process*	*Machine hours per setup*		
A	1	2.4	0.6	1.2
A	2	1.8	—	1.8
A	3	1.2	1.8	1.2
B	1	3.0	1.2	2.4
B	2	0.6	3.0	1.2
B	3	—	3.6	1.2
C	1	2.4	1.8	3.0
C	2	1.2	1.2	1.2
C	3	—	2.4	2.4

weekly machine availabilities and then develop a program based on the assumption that any program would involve exactly six hours total setup time on each type of machine. We can subsequently adjust for the number and kind of setups actually called for by the program.

profitable use which can be made of the available capacity after fulfilling contractual obligations is to produce Product C.

Checking to see how much setup time is actually implied by this program, we discover that it exceeds the six-hour estimate on all three types of

TABLE 10

MOST PROFITABLE USE OF CAPACITY ASSUMING SIX HOURS SETUP PER MACHINE

A — Program Based on Six Hours Setup per Machine

Machine type		*I*	*II*	*III*	*Units produced*
Product	*Process*	*Productive machine hours*			
A	1	18.4	18.4	18.4	92
A	3	4.8	0.8	0.8	8
B	1	40.0	60.0	80.0	200
C	1	48.8	24.4	170.8	244
C	3	—	120.0	30.0	150
	Total	112.0	223.6	300.0	

B — Actual Setup Times Implied by Program

Machine type		*I*	*II*	*III*
Product	*Process*	*Hours of setup time*		
A	1	2.4	0.6	1.2
A	3	1.2	1.8	1.2
B	1	3.0	1.2	2.4
C	1	2.4	1.8	3.0
C	3	—	2.4	2.4
	Total	9.0	7.8	10.2

Table 10 shows the program which would be the most profitable *if this assumption concerning setup were true.* It calls for the production of only the required 100 units per week of Product A and 200 units of B, but it calls for 394 units of Product C instead of just the required 300. In other words, the calculation indicates that the most

machine (see the totals shown in the table under B, Table 10). We could adjust for this by simply reducing the available machine hours accordingly and then recalculating the program, but examination of the program of Table 10 brings to light another fact of which we ought also to take account. This is the fact that only 8 units per week of

Product A are to be manufactured by Process 3.

Since these are bottleneck machines, we do not really need a cost calculation to decide that it is wasteful to tie them up in setup for this almost negligible amount of production. (This decision can be checked, as will be shown shortly.) Therefore we eliminate Process 3 for Product A before adjusting the available machine hours for the amount of setup time actually required, and then recalculate the program, again excluding the unwanted process. One of the more useful features of linear programing is the fact that the calculation need not be purely mechanical, but can always be controlled to agree with common sense.

The resulting revised program is shown in Table 11, together with some related cost information which corresponds to the "row values" and "column values" of the ketchup problems. This information will be discussed more fully in Part III of this article. For the moment we may observe that it confirms our decision to reject Process 3 for Product A. Use of this process for 8 units would save running time worth \$51.20 ($8 \times \$6.40$) but would cost nearly \$100 in setup (1.8 hours on a Type II machine worth \$27.80 per hour plus 1.2 hours on a Type III machine worth \$38.80 per hour).

We could at this point ask whether it might also be better to use only a single process for Product C. Common sense tells us, however, that the production of Product C by each of the two methods is large enough to make setup cost negligible, and again this can be confirmed by analysis of the by-product cost information and other data on the worksheets underlying Table 11. However, the argument is a little more complex than the one concerning Process 3 for Product A and will not be given here.

Features of Program The final program still calls for only the required amounts of Products A and B; proper choice of processes for all products makes it possible to produce 88 units per week of Product C above the minimum requirements. This figure of 88 units is not greatly different from the 94 units shown in the first-approximation program (Table 10). That program, despite the rough-and-ready assumption on which it was based, proved in fact to be a very good guide to the proper use of the available capacity, and only minor refinements were required to make it into the genuinely most profitable program. A more complex problem might, of course, call for several successive approximations instead of just two as in this simple case.

One significant feature of the final program is the fact that it calls for a certain amount of idle time on machines of Type I. Any program which used this type of machine fully would produce *less* profit than the program of Table 11. In one actual application of mathematical programing to a machine shop, a result of exactly this sort proved to be of very considerable practical importance. Without some kind of provable justification, personnel were extremely hesitant to include idle time in the program when management was pressing for all possible production. There is a real danger under such con-

TABLE 11

MOST PROFITABLE USE OF AVAILABLE CAPACITY

A — Revised Program Based on Actual Setup Requirements

Machine type			*I*	*II*	*III*	*Units*
Product	*Process*		*Machine hours*			*produced*
A	1	setup	2.4	0.6	1.2	100
		run	20.0	20.0	20.0	
B	1	setup	3.0	1.2	2.4	200
		run	40.0	60.0	80.0	
C	1	setup	2.4	1.8	3.0	238
		run	47.6	23.8	166.6	
C	3	setup	—	2.4	2.4	150
		run	—	120.2	30.0	
	Idle time		2.6	—	—	
	Total		118.0	230.0	305.6 *	

B — Additional Margin Which Would Be Made Possible by One Additional Machine Hour

Machine type	*I*	*II*	*III*
Margin	—	$27.80	$38.80

C — Loss of Margin Which Would Result from Production of One Unit by Processes Other than Those Selected †

Product	*Process*		
	1	*2*	*3*
A	—	$(1.70) ‡	$(6.40) ‡
B	—	10.00	20.80
C	—	2.20	—

D — Loss of Margin Which Would Result from Production of One Extra Unit of Product Other than Product C

Product	*Loss*
A	$3.30
B	3.00

* Discrepancy from 306.0 due to rounding of figures.

† This table gives the loss which would arise from the running time of the process in question. The loss due to setting up for the additional process can be calculated from the value of one machine hour shown in the previous table.

‡ Minus quantity.

ditions that personnel will produce a program less efficient than is possible simply because they concentrate their efforts on discovering a program which uses all machines 100% of the time.

LOWEST COST PRODUCTION

The last few examples have involved the problem of getting out the most profitable production when a company can produce less than it can sell. Mathematical programing can also be of value when the problem is one of getting out the required production at the lowest possible cost. Here is an interesting example:

> One of the large meat packers is currently using linear programing to find the least expensive way of producing a poultry feed with all the required nutritive values. All that is needed to solve such a problem is: a list of the essential nutrients (minerals, proteins, and so forth) with the amount of each which should be contained in a pound of feed; a list of the possible materials which could be used to produce the feed, with the price of each; and a table showing the amount of each nutrient contained in a pound of each possible constituent for the feed.[6]

This problem is obviously very similar to the avgas problem discussed above, except that here the object of the program is to supply a fixed output at lowest cost rather than to choose the output which will maximize revenue.

Exactly the same kind of problem can arise when there is more than a single end product involved. For instance, the manager of a refinery might be faced with this kind of problem:

> Suppose that instead of having inadequate supplies, this manager has ample capacity to make all he can sell. As we have seen, each of the products which he sells can be blended in a variety of ways from intermediate products such as alkylates and catalytic-cracked gasolines, and each of these intermediate products can be produced out of various crudes in various proportions. The manager of the refinery must decide which crudes to buy and how they should be refined so as to produce the required end products at the lowest possible cost.

Charnes, Cooper, and Mellon have shown that it is possible to use linear programing to solve a still more complex problem than this, bringing in, for example, the possibility of using imported as well as domestic crudes, and considering even such factors as taxes, customs duties, and the cost differences between chartered and company-owned tankers.[7]

Programing can also assist in cost reduction in a machine shop when there is sufficient capacity to produce all that can be sold of every product; it can indicate how to produce each product by the most economical process. All that is required for a programing problem to exist is that the capacity of the company's *best* or most economical machines of a given type—for example,

[6] The use of mathematical programing in connection with a variety of problems in farm economics is described in a number of articles in the *Journal of Farm Economics*, 1951, p. 299; 1953, pp. 471 and 823; 1954, p. 78.

[7] A. Charnes, W. W. Cooper, and B. Mellon, "A Model for Programming and Sensitivity Analysis in an Integrated Oil Company," circulated in mimeographed form by the Carnegie Institute of Technology, and to be printed in a forthcoming issue of *Econometrica*.

its highest-speed screw machine—be less than sufficient for the entire production requirements. To illustrate:

Suppose that a manufacturer wants to produce specified quantities of five different screw-machine parts, A through E, and has available three different screw machines, I, II, and III. Any of the machines can produce any of the parts, but a weekly basis, though we shall assume that management can make each part in long runs and thereby reduce setup cost to a point where it may be neglected in determining the program. Setup, maintenance, and repairs we shall assume to be performed on Saturdays, and therefore we take each machine as being available 40 hours per week.

The lowest-cost program which will accomplish the required production is shown in Table 13 together with the usual by-product cost information. As previously stated, the production shown in the exhibit is in terms of weekly *averages;* the actual length of individual runs can be determined subsequently, in the usual way in which economic lot sizes are determined.

TABLE 12

PRODUCTION RATES, REQUIREMENTS, AND COSTS

Machine	*I*	*II*	*III*	*Average weekly production (units)*
Part	*Per-unit machine time (minutes)*			
A	0.2	0.4	0.5	4,000
B	0.1	0.3	0.5	9,000
C	0.2	0.2	0.4	7,000
D	0.1	0.3	0.3	9,000
E	0.2	0.3	0.5	4,000
	Variable operating cost (per hour)			
	$12	$9	$9	

the rates of operation are different, as shown by the per-unit times in Table 12. If Machine II were slower than Machine I by the same percentage on all parts, and the same were true of Machine III, this problem would not require much thought for its solution, but when the inferiority of a machine depends on the particular part, linear programing is of use.

The hourly variable cost (direct labor, power, repair and maintenance, etc.) of operating each machine is shown in the exhibit, since the machines are not *all* bottlenecks and the whole point of the problem is to avoid operating costs insofar as possible. The exhibit also gives the required average production of each part on

PART III. COST AND PROFIT INFORMATION

Determination of the most profitable program under a particular set of circumstances is by no means the only advantage which management can de-

rive from the intelligent application of mathematical programing. In many situations the technique will be of equal or even greater value as the only practical way of obtaining certain cost and profit information that is essential

A was leading to a reduction of nearly $80,000 a year in profits, far more than had been believed. Now, "cost" in this sense—the difference between the profit which results from one course of action and the profit which would result from

TABLE 13

LOWEST-COST PROGRAM AND BY-PRODUCT COST INFORMATION

A — Lowest-Cost Machine Assignments

	First Alternative Program			*Second Alternative Program*		
Machine	*I*	*II*	*III*	*I*	*II*	*III*
Part	*Average weekly minutes*					
A	600		500	467		833
B	900			900		
C		1,400			1,400	
D	900			900		
E		1,000	333	133	1,000	
Idle time			1,567			1,567
Total	2,400	2,400	2,400	2,400	2,400	2,400

B — Cost of One Additional Unit of Product

Part	*A*	*B*	*C*	*D*	*E*
Cost	$0.0750	$0.0375	$0.0500	$0.0375	$0.0750

C — Value of One Additional Machine Hour

Machine	*I*	*II*	*III*
Value	$10.50	$7.50	$0.00

for sound decisions on both short-run and long-run problems of many kinds.

NEED FOR PROGRAMING

What kind of cost information will mathematical programing provide? The gasoline blending case described in Part II of this article is a good example.

In that instance the management learned that the manufacture of Avgas another course of action—is obviously a completely different thing from cost in the accounting sense. Information regarding this kind of cost cannot be provided by ordinary accounting procedures. In fact, mathematical programing is the only way to get it quickly and accurately when there are many possible combinations of the various factors involved.

Costs for Decision Making In some situations the need for looking at the effect of a proposed action on over-all profits rather than at its accounting cost or profit is perfectly clear. In our gasoline blending case, management knew very well that money was being lost by the production of Avgas A even though the accounts showed a profit; it was only the extent of the loss that was unknown. In other situations, by contrast, accounting cost is really misleading in arriving at a sound decision, and it is easy to overlook this fact. An example should help make this point clear:

It would seem to be plain common sense that the cost of freight to a particular warehouse is simply the freight bill which is paid on shipments to that warehouse. But management will do well to think twice before acting on the basis of this "common-sense" view.

Suppose that the sales manager of the company whose shipping program is given earlier in Table 2 finds that it is becoming very difficult and expensive to sell the supply allocated to Warehouse E, whereas sales could easily be increased at Warehouse T. Selling price is the same at both localities and, because of competition, cannot readily be changed. On inquiry the sales manager finds that Warehouse E is being supplied at a freight cost of 23 cents per cwt., whereas freight to Warehouse T is only 6 cents per cwt. He proposes, therefore, that supplies and sales be diverted from E to T, thus increasing the company's profits by the freight saving of 17 cents per cwt. as well as reducing the cost of advertising and other selling expense.

The traffic manager will probably counter that the two warehouses are not being supplied from the same factory, and that if the supplies now being sent from Factory II to Warehouse E are shipped to Warehouse T instead, freight costs will not fall to 6 cents per cwt., but will increase from the present 23 cents to 54 cents, making a loss of 31 cents per cwt.

Actually, neither of the two would be right. In the event that supplies are diverted from Warehouse E to Warehouse T, there will in fact be an extra freight cost rather than a saving. But *if the change is properly programed* (the supplies formerly sent from II to E should be sent to Q, which can then take less from XII, which in turn can then supply the additional amount to T), then the extra cost will be only 14 cents per cwt. It is this cost which management should compare with the estimated extra cost of selling at Warehouse E.

The example just cited and the gasoline blending case are typical of the way in which mathematical programing can be used to calculate the cost or profit which results or will result from a management decision. Generally speaking, any program is determined in such a way as to produce the greatest possible profits under a certain set of fixed conditions. If management wishes to consider a change in any of these conditions, a new program can be computed and profits under the two sets of conditions can then be compared.

Available Figures In a some cases it is not even necessary to compute a new program to find the cost or profit which applies to a proposed decision. The computation of the original program itself yields as a free by-product the cost or profit which will result from certain changes in the conditions underlying the program, *provided that these changes are not too great in ex-*

tent. In the jargon of the economists, these by-product figures are "marginal" cost or profit rates. To illustrate:

For diversion of sales from Warehouse E to Warehouse T, the marginal cost is given immediately by comparison of the "row values" shown in Table 2 for the two warehouses. The value for E is 28 cents per cwt., the value for T is 42 cents, and the extra cost is therefore 14 cents per cwt. (42–28). We can be sure at once that this will be the extra cost if only a single cwt. is diverted from one warehouse to the other, but in order to find the cost of a larger diversion we must study the program itself. If we do so, we will find that the marginal rate will hold in this case even if the entire supply now allocated to E is diverted to T. If, on the contrary, we were considering diversion from Warehouse G to T, we would find that the marginal rate of 15 cents (42–27) would apply only to the first 180 cwt.

The "column values" of Table 2 give similar information concerning the cost or saving which will result from shifting production from one plant to another. If production is increased at Factory V and decreased at Factory VI, there will be a saving of 13 cents per cwt. (−38–[−51]) up to a certain limit, and study of the program shows that this limit is again 180 cwt.

The costs shown in Tables 11 and 13 are marginal rates of this same sort. In fact, such information could have been given in connection with all the programs developed in this article.

Probably the most important use of the marginal rates is that they immediately give a *minimum* figure for the cost of a change which *reduces* profits, or a *maximum* figure for the profitability of a change which *increases* profits. For example, when the program of Table 11 shows that an additional hour on a machine of Type III is worth $38.80, we can be sure that ten additional hours will be worth no more than $388, although they may be worth less. Inspection of the marginal costs can thus be of practical value in limiting the range of alternatives which are worth further investigation.

USES OF INFORMATION

Now let us turn to consider a number of examples of particular kinds of cost and profit information which can be obtained by mathematical programing and which will be of use in making management decisions.

Product Cost The gasoline blending case was as good an illustration as possible of the use of mathematical programing to find the true profitability of a particular product, but the technology of gasoline blending is so complex that it is not easy to see why the answer comes out as it does. Since it is difficult to make intelligent use of a technique without really understanding how it operates, let us look briefly at a much simpler example of the same kind of problem:

In the first case involving the assignment of machine tools in Part II, there was idle capacity available after meeting the contractual commitments (see Table 13). Suppose that, after this schedule has been worked out, a customer places an order for an additional 1,000 units of screw-machine Part D. What will be the cost of filling this order?

Machine III is the only machine with idle capacity; and if the additional quan-

tity of Part D is made on that machine, it will cost $75 (500 minutes at $9 per hour). The most economical course of action, however, is to produce the additional 1,000 units of D on Machine I, obtaining the required 100 minutes by taking 500 units of Part A off this machine and putting them on Machine III. If this is done, the accounting cost of the 1,000 units of D will be only $20 (100 minutes at $12 per hour), but the actual addition to total cost will be $37.50 (250 minutes at $9 per hour to make the 500 units of A on Machine III). Thus the true cost of the additional 1,000 units of D will be $0.0375 each, the value shown in Table 13. Any price above the sum of this figure and the material cost of the part will make a contribution to fixed overhead.

Most Profitable Customers The example of the diversion of sales from Warehouse E to Warehouse T previously discussed shows how programming can be used to determine which customers are the most profitable in a situation where the only difference among customers lies in the cost of freight. The question would be no harder to answer if some customers were supplied from plants with higher production costs than others. Actually, of course, there is very little difference between determining the profitability of a product and the profitability of a customer.

Marketing Policy Cost and profit information calculated by mathematical programing can be of use to management in deciding what products to make, what prices to set, and where to expend selling effort. We wish to emphasize, however, that we are not proposing that management should build its entire marketing program on the basis of short-run profit considerations. Programing provides information; it it does *not* provide answers to policy questions.

On learning that certain products or certain customers are relatively unprofitable under present conditions, it is up to management first of all to decide whether the situation is temporary or likely to continue for some time to come. This means that management should forecast future costs and future sales potentials under a variety of reasonable assumptions, and then calculate the profitability of the various products or markets under various combinations of these assumptions. It is here that mathematical programing will make its real contribution, since it is only when such calculations can be easily and cheaply carried out that management can afford to investigate a wide range of assumptions.

After such calculations have been made, management can decide to change prices, refuse certain orders, accept them at a short-run loss, or install new capacity of such a kind and at such places that the products or markets in question will become profitable.

Cost of Improvements Another kind of cost which it is often important to know is the cost of an improvement in the quality of product or service rendered to the customer. A similar problem arises when it is necessary to decide whether improved materials acquired at higher cost will increase revenues or reduce other costs sufficiently to justify their higher cost. Here are some illustrative cases:

1. *Cost of quick delivery*—According to the shipping program of Table 2, Warehouse M is to be supplied partly from Factory II at a cost of 40 cents per cwt. and partly from Factory IV at 21 cents per cwt. Suppose that stocks are low at this warehouse and that the manager would like to obtain some supplies quickly from the nearest source, Factory V. Since this is the nearest plant, the freight rate to Warehouse M, 10 cents per cwt., is naturally lower than the rates from the factories currently supplying the warehouse; but use of this shorter route will necessarily result in an *increase* in total cost, since the program as it stands gives the lowest possible total cost.

Programing shows immediately that the extra cost will be 16 cents per cwt. for the first 140 cwt. shipped to M from Factory V. The higher cost applying to additional quantities could be readily calculated if it were needed.

2. *Choice of process in a machine shop*—In the case of the machine shop with limited total capacity, Table 11 showed that the most profitable course of action was to produce Product B by the use of Process 1. Suppose that while an adequate product results from this process, a better quality would result from the use of Process 3. Would it be worth using this process in order to increase customer satisfaction, or could the price be increased sufficiently to recover a part of the additional cost?

The program of Table 11 shows immediately that the extra cost resulting from the use of Process 3 for Product B will be at least $20.80 per unit. The cost arises because use of this process instead of Process 1 takes up capacity which is being used for the production of Product C, each unit of which produces a "margin" of $30 per unit. Up to 128 units of B can be made by Process 3 instead of Process 1 at the cost of $20.80 per unit. If 128 units are made, the entire capacity of the shop will be used up in producing the contractual commitments for the three products, and further use of Process 3 for Product B will be impossible.

3. *Cost of antiknock rating*—In the gasoline refinery studied by Charnes, Cooper, and Mellon, antiknock ratings (PN's) were specified for Avgas B and Avgas C for both rich and lean mixture. During the study an interesting question was raised as to the additional cost entailed by the rich-mixture specification. It was found to amount to over $1,000 per day. In other words, profits could have been increased by that amount if only a lean-mixture rating had been required in the products. A little further calculation with their data produced the equally interesting result that the lean-mixture requirement on these two fuels was costing nothing; satisfaction of the rich-mixture requirement automatically produced oversatisfaction of the lean-mixture requirement.

4. *Value of improved materials*—Engineers of this same refinery suggested that if the volatility of the straight-run gasoline being used in blending could be reduced, it would be possible to produce a product mix with a considerably higher market value. Again, programing provided significant and accurate information. It was able to show that if the RVP of this stock could be reduced by one unit, from 4.0 to 3.0, the market value of the products could be increased by $84 per day. Thus, if the improved stock could be produced at an additional cost smaller than this, it would pay to do so; otherwise it would not.

Capital Investments Some of the most important decisions that management has to make are those which involve the choice of the most profitable ways in which to invest new capital. The choice is usually made by compar-

ing the cost of each proposed investment with the increase of income that it will produce. When several of the proposed investments are for use in the *same* productive process, and when this process produces a variety of different products, it may be extremely difficult to determine the additional income that will result from any one investment or from any combination of investments without the use of a systematic computing technique.

Machine Tools. Consider, for example, the machine-shop case described in Part II in which sales were limited by machine capacity. Under the program of Table 11, all machines of Type II and Type III are loaded to capacity; and while there is idle time on machines of Type I, it is very small in amount and actually exists only because it was unprofitable to set up to produce just 8 units per week of Product A by Process 3. Under these conditions what would be the return on an investment in an additional machine of one of the three types? It will be enough to work out the answer to this question for just one of the three types as an example, assuming that management has forecast that present demand and present costs and prices will remain unchanged in the future:

Suppose that if the shop acquires one additional machine of Type III, it would be available for 38 hours per week (one shift with allowance for down time). We simply calculate a new program for the same conditions as shown in Table 9, except that we increase the available time on machines of Type III from 300 to 338 hours. The resulting program shows a $960 per-week increase in "margin"—selling price less all costs of production except the costs on the bottleneck machines. (To find the additional *income* produced by the new machine, we would have to subtract the labor and overhead costs of operating the machine and the depreciation and other costs of owning it.)

The result is due to the fact that the additional machine will make it possible to produce 32 additional units of Product C per week. Note that the $960 margin on 38 hours of use amounts to only $25.30 per hour, considerably less than the $38.80 shown in Table 11. As more time is made available on machines of Type III, the bottleneck on this type becomes relatively less important and the bottlenecks on the other two types become relatively more important.

Raw Materials. Without actually working out examples, we can point to either the gasoline refinery or the hypothetical case on the selective use of raw materials (both in Part II) as two other situations where the profitability of investment would be very difficult to calculate without the use of mathematical programing. The refinery problem discussed above involved only the most profitable way in which to blend *existing* supplies of materials. Mathematical programing would readily show the additional sales revenue which could be obtained (at present prices) if the refinery were to enlarge its facilities for production of one or more of the blending stocks.

In the case on selective use of raw materials, the materials had to be purchased in the market; and, as shown in Table 6, it proved unprofitable to produce Product D because of the limited supplies of materials available at

normal prices. Programing could readily show how much the company could afford to invest in a source of raw materials in order to obtain them at more reasonable cost.

Programing and Forecasts. In the case of investment decisions even more than in the case of the other types of decisions previously discussed, the relevant data are not so much the facts of the immediate present as they are forecasts of conditions which will prevail in the future. An investment decision cannot be made rationally unless it is possible to explore its profitability under a variety of assumptions about future costs and markets.

It is already difficult enough to make the necessary forecasts; without the use of a systematic technique for calculation, full exploration of their implications is virtually impossible because of time, trouble, and expense. It is for this reason that it seems likely that mathematical programing may be of even greater value to management in the field of planning than in the field of immediate operating decisions.

As in the case of its other applications, however, mathematical programing is not a cure-all. Management can use it to great advantage in planning and policy making, but executives must first understand it correctly and be able to use it intelligently in combination with the other tools of forecasting and planning. The fate of mathematical programing, in other words, lies today in management's hands. The scientists, the inventors, have done their work; it is now up to the users.

APPENDIX. DIRECTIONS FOR SOLVING PROBLEMS BY A USEFUL SHORT PROCEDURE

There are several alternate procedures available for solving problems of linear programing. One of these will work in all cases but takes a long time to carry out—the "general procedure," which is discussed toward the end of this appendix. The others are relatively quick, but will work only in certain cases—e.g., the "profit-preference procedure" and the "transportation-problem procedure."

A very restricted class of problems can be solved by hand with remarkable ease through the use of the "profit-preference procedure." A good example of its use is the scheduling of two classes of machine tools which formed a bottleneck in the operations of one actual company. The example has been published, with clear instructions for carrying out the procedure.[8]

By far the most frequently useful of the shorter procedures is the one known as the "transportation-problem procedure." [9] As pointed out in the preceding text, it got this name because it was developed to determine lowest-cost

[8] See A. Charnes, W. W. Cooper, and D. Farr, "Linear Programming and Profit Preference Scheduling for a Manufacturing Firm," *Journal of the Operations Research Society of America I*, May 1953, pp. 114–129. (The reader should be warned that errors have crept into Tables III and IV of this publication.) The technique is similar to the one used by SKF: cf. above.

[9] This procedure was developed by G. B. Dantzig: see T. C. Koopmans, *Activity Analysis of Production and Allocation* (New York, John Wiley & Sons, Inc., 1951), pp. 359–373.

shipping programs, but it can be used for problems not involving transportation (just as certain problems involving transportation cannot be solved by it). Because of its simplicity, we shall give full directions for its use, first working through a simple example and then giving some suggestions for reducing more complex problems to such a form that they can be solved in the same way.

it is understood; some suggestions for doing that will be given.

Table A gives the data for the problem: the freight rates from each plant to each warehouse, the capacity of each plant, and the requirements of each warehouse. Now let us go through the various steps of the solution.

Getting a Starting Program. We first

TABLE A

RATES, REQUIREMENTS, AND CAPACITIES

Factory	*I*	*II*	*III*	*Warehouse*
	Freight rates (dollars per ton)			*requirements (tons)*
Warehouse A	1.05	.90	2.00	35
B	2.30	1.40	1.40	10
C	1.80	1.00	1.20	35
D	1.00	1.75	1.10	25
Factory capacity (tons)	5	60	40	105

TRANSPORTATION-PROBLEM PROCEDURE

Our example consists of assigning the production of three plants to fill the requirements of four warehouses in such a way that the total cost of freight will be at a minimum. This example involves so few variables that it could be solved far more quickly by common sense than by the use of a formal procedure. The example is adequate, nevertheless, to explain the procedure, and the procedure can then be used to solve much larger problems that would be extremely difficult to solve by common sense. Furthermore, the procedure itself can be considerably short-cut once

get a shipping program which satisfies the fixed requirements and capacities, regardless of cost, by the following procedure. Take Factory I and assign its 5 tons of capacity to Warehouse A. Fill the remaining 30 tons of this warehouse's requirements from Factory II. Then use 10 more tons of Factory II's capacity to satisfy Warehouse B, and assign its remaining 20 tons in partial satisfaction of Warehouse C. Complete C's requirements from Factory III, and use the remainder of III's capacity to satisfy Warehouse D. This produces the starting program of Table B. The procedure could obviously be used to assign warehouses to factories in a problem of any size.

TABLE B

INITIAL PROGRAM OF SHIPMENTS (TONS)

Factory	*I*	*II*	*III*	*Total*
Warehouse A	5	30		35
B		10		10
C		20	15	35
D			25	25
Total	5	60	40	105

A starting program can be based on a guess at the best solution rather than on the "blind" procedure described in the text; and if the guess is any good at all, subsequent calculation will be materially reduced. Start with any factory at all and use its capacity to fill the requirements of those warehouses which it seems most economical to assign to this factory. When that factory's capacity has been used up, take any other factory; first use its capacity to complete the requirements of the warehouse which was left only partially satisfied at the end of the previous step, and then go on to fill any other warehouses which it seems sensible to assign to the second factory.

The only rule which should not be neglected is to finish filling the requirements of one warehouse before going on to a new one. If the number of plants is greater than the number of warehouses, it is perfectly legitimate, however, to reverse the procedure. Start by assigning one warehouse to a series of plants, and, when the warehouse's requirements are filled, take the next warehouse, use it to absorb the leftover capacity of the last factory previously used, and then go on to new factories.

The easiest way to do the work is on paper ruled into squares; and in the following discussion reference is made to locations in the tables as "squares"; for example, the number located in Row B and Column III is said to be in Square B III.

Row Values and Column Values. Next build up a "cost table" by the following procedure:

(1) Fill in the actual freight rates, taken from Table A, for those routes which are actually in use in Table B. This produces Table C except for the "row values" and "column values."

(2) Fill in the "row values" and "column values" shown in Table C. To do this, assign an arbitrary row value to Row A; we have chosen .00 for this value, but it might have been anything. Now under every square of Row A which contains a rate, assign a column value (positive or negative) such that the sum of the row and column values equals the value in the table. In Column 1 we put a column value of 1.05, since 1.05 + .00 gives the value 1.05 found in Square A I; in Column II we put a value of .90, since .90 + .00 gives the .90 in Square A II.

(3) We have now assigned all the

column values which we can assign on the basis of the row value for Row A. We must next assign additional row values on the basis of these column values. We therefore look for rows with no row value but containing rates in squares for which column values exist. We observe that Rows B and C both have rates in Column II, which has a

ing row and column values can always be extended to fill in the row and column values for any cost table provided that "degeneracy" is not present in the corresponding route table. Degeneracy will be explained and a method of dealing with it will be described subsequently. In the absence of degeneracy, inability to complete the row and col-

TABLE C

RATES FOR ROUTES USED IN TABLE B (DOLLARS PER TON)

Factory	*I*	*II*	*III*	*Row value*
Warehouse A	1.05	.90		.00
B		1.40		.50
C		1.00	1.20	.10
D			1.10	.00
Column value	1.05	.90	1.10	

column value of .90. The row value for Row B must be set at .50, since .90 + .50 = 1.40, which is the rate in B II. By the same reasoning, we arrive at .10 as the row value for Row C.

(4) No further row values can be assigned, so we go back to assigning column values by looking for rates which now have a row value but no column value. We observe that there is a 1.20 in Square C III, which has a row value of .10 but no column value. The column value must be 1.10 in order to have 1.10 + .10 = 1.20.

(5) Finally, we assign the one missing row value. In Row D there is 1.10 in Square D III, with a column value of 1.10 and no row value. The row value must be .00 if the total of the row and column values is to equal the value in the square.

This procedure of alternately assign-

umn values, or the existence of contradictory evidence on row and column values, indicates that an error has been made either in drawing up the table of routes (Table B) or in putting down in the cost table (Table C) the rates which correspond to the routes in Table B. On the other hand, it is not essential to derive the row values in the order A, B, C, D and the column values in the order I, II, III; they may be derived in any order that is possible.

The Cost Table. We now proceed to make Table C into a complete cost table, Table D, by filling in all the blank squares with the total of the appropriate row and column values. For example, the 1.55 in Square B I is the total of the row value for Row B (.50) and the column value for Column I (1.05). The figures thus derived are

TABLE D

COSTS FOR ROUTES USED IN TABLE B (DOLLARS PER TON)

Factory	*I*	*II*	*III*	*Row value*
Warehouse A	**1.05**	**.90**	1.10	.00
B	1.55	**1.40**	1.60	.50
C	1.15	**1.00**	**1.20**	.10
D	1.05	.90	**1.10**	.00
Column value	1.05	.90	1.10	

shown in Table D in lightface type, whereas the figures taken from Table C and corresponding to routes actually in use (in Table B) are shown in boldface type. (In practice, the cost table can be made up directly without actually filling in the row and column values.)

Revising the Program. We now have a complete set of tables: a rate table, a route table, and a "cost" table. We proceed to look for the best change to make in the route table in order to reduce the cost of freight. To find this change, we compare the cost table, Table D, with the rate table, Table A, looking for the square where the figure in Table D is *larger* than the corresponding figure in Table A by the greatest difference. This is Square B III. The fact that Table D shows 1.60 while Table A shows 1.40 tells us (for reasons to be explained later) that if we make shipments from Factory III to Warehouse B, and make the proper adjustments in the rest of our program, we shall save 20 cents for every ton we can ship along this new route.

The next problem is to find out what adjustments will have to be made in the rest of the program and, thereby, to find out how much we *can* ship along the new route from III to B. To do this, we construct Table E by first copying Table B (in actual practice there would be no need to copy the table) and then going through the following procedure.

(1) In the Square B III write $+x$: this is the as yet unknown amount which will be shipped over the new route from III to B. We have now over-

TABLE E

CHANGES TO BE MADE IN ROUTES OF TABLE B (TONS)

Factory	*I*	*II*	*III*	*Total*
Warehouse A	5*	30*		35
B		$10 - x$	$+ x$	10
C		$20 + x$	$15 - x$	35
D			25*	25
Total	5	60	40	105

loaded the capacity of Factory III by the amount x, and must therefore decrease by x the amount which III is to supply to some other warehouse. When this is done, it will be necessary to supply this warehouse from some other factory, and so on.

(2) To locate the factories and warehouses which will *not* be affected, look through Table E and put a star beside any number which is the only number in *either* its row *or* its column, but remember that the x in B III counts as a number. This leads to putting a star beside the 5 in A I and the 25 in D III. Considering the starred numbers as nonexistent, look through the table again and put a star beside any numbers which are now left alone in their row or column owing to the elimination of the starred numbers in the previous step. This leads to putting a star beside the 30 in A II, since with the 5 in A I starred, A II is alone in its row.

Now look through the table again for additional numbers which have been left alone in their row or column. In this case we can find none, so the operation is complete; otherwise, we would continue eliminating until no more isolated numbers could be found.

(3) Having completed the foregoing procedure, we now make all required adjustments by changing the amount to be shipped along those routes which have *not* been eliminated by a star. (Once a little experience has been gained, the routes affected by a change can easily be found without first starring the routes not affected.) The +x in B III overloads Factory III, so write −x beside the 15 in C III. Warehouse C is now short by x, so write +x beside the 20 in C II. Factory II is now overloaded, so write −x beside the 10 in B II. This last −x balances the +x in Row B with which we started, so that the effect of using the new route has been completely adjusted for throughout the program.

(4) Since we shall save 20 cents for every ton we ship along the new route from III to B, we wish to divert as much tonnage as possible to this route. We therefore look at all the squares in which we have written −x and discover that the smallest number with −x beside it is the 10 in B II. This is the limit to the diversion, and therefore the value for the unknown x. We now produce Table F by subtracting 10 in Table E wherever −x was written and

TABLE F

FIRST REVISED PROGRAM OF SHIPMENTS (TONS)

Factory	*I*	*II*	*III*	*Total*
Warehouse A	5	30		35
B			10	10
C		30	5	35
D			25	25
Total	5	60	40	105

TABLE G

COSTS FOR ROUTES USED IN TABLE F (DOLLARS PER TON)

Factory	*I*	*II*	*III*	*Row value*
Warehouse A	**1.05**	**.90**	1.10	.00
B	1.35	1.20	**1.40**	.30
C	1.15	**1.00**	**1.20**	.10
D	1.05	.90	**1.10**	.00
Column value	1.05	.90	1.10	

adding 10 wherever +x was written. This is our first revised program of shipments. By multiplying the shipments along each route by the rate for that route, the reader can check that the reduction in total freight cost has in fact been 20 cents per ton times the 10 tons diverted to the new route.

Repeating the Process. The rest of the solution proceeds by mere repetition of the process already followed for the first improvement in the program. We build up a new cost table, Table G, by first copying from Table A the rates for the routes used in Table F (these rates are shown in boldface type in Table G), then calculating the row and column values, and then filling in the other squares (lightface type). We next compare Table G with Table A square by square and find that the square with the largest difference in favor of G is D I (1.05 against 1.00). We therefore put +x in D I of Table H, remove the "isolated" squares with stars, and then follow around a circuit with +x and −x as indicated. The square with the smallest number with a −x beside it is A I, with a value of 5, and we therefore add or subtract 5 as indicated by +x or −x to produce Table J.

From Table J we make up a new cost table, Table K. Comparing Table K with Table A, we find that *every* lightface figure in Table K is smaller than the corresponding figure in Table A. There is no further improvement that can be made; in fact, any change made

TABLE H

CHANGES TO BE MADE IN TABLE F (TONS)

Factory	*I*	*II*	*III*	*Total*
Warehouse A	5 − x	30 + x		35
B			10*	10
C		30 − x	5 + x	35
D	+ x		25 − x	25
Total	5	60	40	105

TABLE J

SECOND REVISED PROGRAM OF SHIPMENTS (TONS)

Factory	*I*	*II*	*III*	*Total*
Warehouse A		35		35
B			10	10
C		25	10	35
D	5		20	25
Total	5	60	40	105

in the program of Table J would result in an *increase* in the cost of freight. Had there been squares where the lightface figure in Table K was just equal to the rate in Table A, this would have indicated a route which could be used without either raising or lowering the total cost of freight.

Why the Procedure Works. To see why this method works, consider the map shown in Chart A. This map corresponds to the shipping program shown in Table B, with a solid line joining every factory to every warehouse where shipments are to be made. Beside each line is shown the tonnage moving along the route together with the freight rate applying to that route according to Table A. The map also shows a dotted line from Factory III to Warehouse B, corresponding to the x which we put in Square B III in Table E.

Now suppose that we ship x tons from Factory III to Warehouse B. Every ton that we ship will cost $1.40, the rate between these two points. But for every ton which B gets from III, one less ton from II will be needed, thereby saving $1.40 of freight. Factory III, on the other hand, cannot now supply both C and D as before, whereas Factory II now has an excess. The simplest solution is to have III ship less to C, thus saving $1.20 per ton, while II makes up the deficit at a freight cost of $1.00 per ton. The net effect is a saving of 20 cents per ton, even though the shipments from III to B cost just as much as the previous shipments from II to B.

TABLE K

COSTS FOR ROUTES USED IN TABLE J (DOLLARS PER TON)

Factory	*I*	*II*	*III*	*Row value*
Warehouse A	1.00	**.90**	1.10	.00
B	1.30	1.20	**1.40**	.30
C	1.10	**1.00**	**1.20**	.10
D	**1.00**	.90	**1.10**	.00
Column value	1.00	.90	1.10	

This saving of 20 cents per ton is exactly the difference between the $1.60 in Square B III of Table D and the $1.40 in the same square of Table A. This is true in general; the lightface figures in a "cost table" show the *net* savings on *other* routes which can be made by readjusting the program if direct shipments are made along the route in question. In other words, the lightface figures show the cost of "not using" a route; the cost of using the route is, of course, simply the freight rate as shown in Table A.

The best possible program has not been reached until there is no unused route for which the cost of "using" is less than the cost of "not using." To be sure, at any stage in the process of arriving at a best program there may be more than one route for which the cost of not using is higher than the cost of using. We have given the rule of making the change by introducing the route for which the difference between the two costs is greatest. This rule is not necessary, but it is commonly believed that use of this rule will *usually* reduce the number of steps required to arrive at the best possible program.

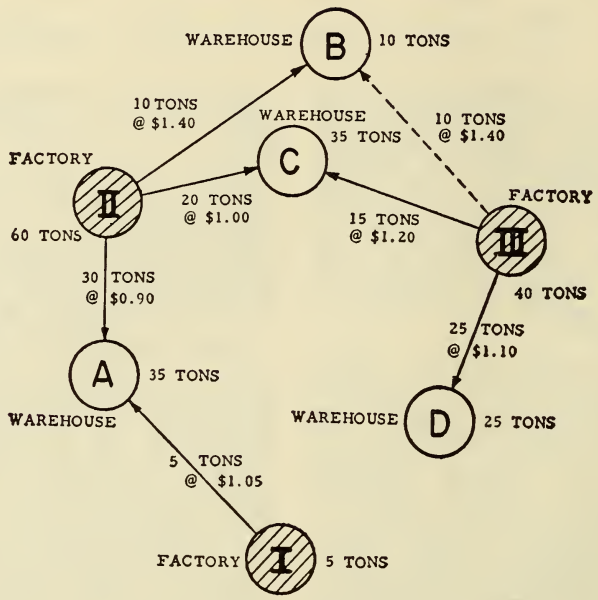


CHART A

MAP OF ROUTES USED IN TABLE B

Any program is a best possible program if there is no unused route for which the cost of using is less than the cost of not using. This is a rather important fact, since it means that a solution can be checked by simply building up the corresponding cost table. There is no need to check over the work

which produced the solution. Furthermore, if there is an error in the solution, it is a waste of time to go back to find it; everything will come out all right if you simply go on making successive changes until the best possible program emerges. This is an additional reason why the transportation-problem procedure is really suited for hand computation while the general procedure is not; there is a reasonably simple check on the accuracy of the final solution obtained by the general procedure, but correction of any errors that may be present is far more difficult.

The map also shows why we arrived at the value 10 for the x in TABLE E. If we make direct shipments from III to B, we must reduce shipments from II to B and from III to C. We cannot reduce either of these below zero. The route from II to B carries the smaller traffic, 10 tons, and therefore 10 tons is the largest amount we can ship from III to B. Table E has $-x$ beside each route that will be reduced as a result of the change, and a $+x$ beside each route that will be increased. The routes which are starred in Table E are the routes which are *not* in the "circuit" III-B-II-C-III.

In some cases adjustments could be made which would give a greater saving per ton or make possible diversion of more tons than will result from the use of the rules given above. It is perfectly permissible to make more general changes in the program at any stage provided that they are made in accordance with the rule given previously for starting the program. On the other hand, such general adjustments are never *necessary,* since it is absolutely certain that the step-at-a-time method described above will ultimately lead to the best possible program.

Coping with Degeneracy. The procedure just described serves to solve any "transportation" problem of any size except when degeneracy appears in a route table at some stage in the solution of the problem.

A route table is degenerate if it can be divided into two or more parts each of which contains a group of factories whose combined capacity exactly satisfies the combined requirements of the warehouses assigned to them. Table L gives an example of such a situation which might have arisen in solving the example we have just worked out. Warehouses A and D exactly use up the capacity of Factory II, while Warehouses B and C exactly use up the capacity of Factories I and III. Under such circumstances the procedure breaks down because it is impossible to build up the cost table corresponding to a degenerate route table; that is, in this instance, the cost table corresponding to Table L.

The following simple device will take care of this difficulty: If the number of plants is smaller than the number of warehouses, divide one unit of shipment by twice the number of plants. (If shipments are to be measured to the tenth of a ton, for example, we divide 1/10 ton, *not* 1 ton, by twice the number of plants.) Take any convenient number which is smaller than this quotient and add it to the capacity of each of the plants; add the same *total* amount to the capacity of any one warehouse. If the number of warehouses is

less than the number of plants, then reverse the rule.

In either case, solve the problem as if the additional quantities were real parts of the requirements and capacities; then when the problem has been solved, round all numbers containing fractions to the nearest unit of shipment. (A route carrying less than one-half unit is rounded to zero.) *The solution thus obtained is not approximate; it is exact.*

TABLE L

PROGRAM OF SHIPMENTS WHICH MIGHT HAVE OCCURRED BEFORE REACHING SOLUTION

Factory	*I*	*II*	*III*	*Total*
Warehouse A		35		35
B	5		5	10
C			35	35
D		25		25
Total	5	60	40	105

WHEN TO USE

In its original application, as illustrated in the example worked through above, the transportation problem consists of assigning a set of sources to a set of destinations in such a way that the total cost of transportation from sources to destinations will be a minimum. The capacity of each individual source and the requirements of each individual destination are fixed in advance, and the total capacity equals the total requirements. A unit of requirements at any destination can be filled by the use of a unit of capacity at any source, and only the cost of freight varies according to which particular source is used.

This can easily be generalized as a problem of assigning a set of *inputs* of any nature whatever to a set of *outputs* of any nature whatever in such a way that the total *cost of conversion* is a minimum. The inputs might be the available supplies of various raw materials, for example, rather than the capacities of various factories, while the outputs might be the quantities produced of various products rather than the quantities of a single product shipped to various warehouses.

There is no real change when the problem is one of maximizing profits rather than minimizing costs. Instead of a "rate table" giving the cost of converting one unit of any input into one unit of any output, we have a "margin table" giving the margin which will be realized by such conversion, the margin being the revenue from selling the unit of output less the variable costs of producing it. The program is developed in exactly the same way as in the example worked through above, except that new "routes" are introduced when the margin from not using the route is less than the margin from using it, rather than when the cost of not using it is higher than the cost of using it.

The formal characteristics which a

problem must have if it is to be solved by the transportation procedure are the following:

(1) One unit of *any* input can be used to produce one unit of *any* output.

(2) The cost or margin which will result from conversion of one unit of a particular input into one unit of a particular output can be expressed by a single figure regardless of the number of units converted.

(3) The quantity of each individual input and output is fixed in advance, and the total of the inputs equals the total of the outputs.

If a problem cannot be put into the form specified by these three characteristics, it cannot be solved by the transportation procedure. However, these are *formal* characteristics, and it is often possible to find devices or tricks which will put a problem into this form even though at first glance it seems quite different. It is impossible to give a complete list of such devices, but we shall describe here the more common ones. . . .

Inputs and Outputs Not Fixed in Advance. In many problems all that we know in advance is how much of a given input is *available* or how much of a given output *could* be sold. We wish the program to determine how much of each it will be profitable to use or make. This violates the third requirement stated above, but the difficulty is easily overcome by the introduction of "dummy" inputs and outputs.

If, for example, total factory capacity exceeds total warehouse requirements, we create a dummy warehouse and treat it exactly as if it were real. The cost or profit which will result from supplying a unit to the dummy warehouse from any factory is set down in the rate table as zero, and the requirements of the dummy warehouse are set equal to the difference between total capacity and total real requirements. That part of any factory's capacity which the final program assigns to the dummy warehouse is capacity which is actually to be left idle.

If total potential output exceeds total available input, we create a dummy input equal to the difference between the two. The cost or margin resulting from supplying a unit of output from the dummy input is set at zero in the cost or margin table; where the final program calls for producing all or a part of some output from the dummy input, that amount of this potential output is not really to be produced at all.

In a case such as that described in Part II under the heading Where to Sell, it is possible that potential inputs may be left unused at the same time that potential outputs are left unfilled. The calls for the use of both a dummy input and a dummy output. Since neither the total amount of real inputs which will be used nor the total amount of real outputs which will be produced is known until the program has been computed, the quantity of the dummy input must be set equal to or greater than the total of the potential real outputs, and the amount of the dummy output must be set equal to or greater than the total of the potential real inputs. With this proviso, the quantities assigned to the dummies are arbitrary, except that the total of the real plus dummy *in*puts must equal the total of

the real plus dummy *out*puts. The final program will show a certain amount of dummy output to be supplied from the dummy input, but this figure has no real meaning whatever and should be disregarded.

Inputs and Outputs at Varying Prices. It may be that a factory can supply a certain amount of product at one cost and an additional amount at a higher cost (for example, by the use of overtime), or that a certain amount of a material can be obtained at one price and additional quantities at higher prices. Similarly it may be possible to sell a certain amount of product at one price and additional amounts only at lower prices. All such cases are handled by treating the input at each cost as a separate input, or the output at each price as a separate output. In this way we can still produce a cost or margin table which shows a single unchanging per-unit cost or margin for converting any particular input into any particular output.

Note that this method will *not* work if the price at which the *entire* output is sold depends on the quantity sold. As pointed out in Part II under the heading Price, Volume, and Profit, this is not a problem of *linear* programing.

Impossible Processes. The first formal requirement set forth above demands that one unit of any output be producible from one unit of any input. In some cases particular input-output combinations may be completely or practically impossible. For example, freight service uniting a particular factory with a particular warehouse may be so poor that management will in no case permit its use, or it may be simply impossible to make a particular product from a particular material. This situation causes no difficulty at all in the solution of the problem, since all we need to do is to assign a fictitious, extremely high "cost" to the conversion of this input into this output. In this way we can be sure that the unwanted process will not appear in the final solution.

Artificial Units. In other problems, the *amount* of output which can be obtained from a unit of input depends on the particular output and input in question. In problems involving the selective use of raw materials, for example, the yield of any material may depend on the product, and the amount of material required for a particular product may depend on which material is used. Usually such problems cannot be solved by the transportation procedure, but in some cases the data can be reduced to such a form that they can.

This was true in the first raw-material problem discussed above. The trick here was to express each output not in terms of the quantity of product but in terms of the amount of Grade I material which would be required to produce it, and to express the inputs of Grade II and Grade III material in terms of the amount of Grade I material which they could replace. This made it necessary, of course, to make corresponding changes in the per-unit purchase cost of Grades II and III material and in all per-unit processing costs. TABLE M shows the form to which Table 5 had to be reduced before computing the program of Table 6.

The reason why the subsequent cases discussed above could not be solved by the transportation procedure should now be clear. If the raw-material problem were changed so that the inferiority in yield of the lower-grade materials varied from product to product, it would no longer be possible to express these inputs in such a way that one unit of any input could produce one unit of any output. In the machine-shop problems, the amount of time on one machine which could be replaced by one hour on another varied according to the product and the process being used. The avgas problem is still more complex, since a single unit of any output is blended from several inputs.

TABLE M

MARGINS, SALES POTENTIALS, AND AVAILABILITIES

Product	*A*	*B*	*C*	*D*	*Quantity available (equivalent tons)*
Material	*Margin per equivalent ton*				
I at $48/ton	$ 17	$ 32	$ 4	$ 17	100
I at $72/ton	(7)*	8	(20)*	(7)*	100
II at $24/ton	19	24	12	14	100
II at $36/ton	1	6	(6)*	(4)*	100
III at $18/ton	15	24	16 †	(5)*	150
III at $24/ton	5	14	6	(15)*	250
Potential sales (equivalent tons)	240	150	240	90	

* Minus quantity.

† *Derivation for Product C and Grade III material at normal price.* As shown by the yield table (Table 5), 2.5 tons of III replace 1.5 tons of I, so that 1 ton of III = .6 equivalent tons. As shown by the same table, 1.5 tons of I are required to produce 1 ton of C, so that 1 ton of C = 1.5 equivalent tons.

Material available: 250 tons, or .6 × 250 = 150 equivalent tons.

Sales potential: 160 tons, or 1.5 × 160 = 240 equivalent tons.

Product price: $135 per ton, or $135/1.5 = $90 per equivalent ton.

Processing cost: $66 per ton of product, or $66/1.5 = $44 per equivalent ton.

Material cost: $18 per ton, or $18/.6 = $30 per equivalent ton.

Margin: $90 (selling price) −$44 (processing cost) −$30 (material cost) =$16 per equivalent ton.

Such are the problems which call for the use of the general procedure.

THE GENERAL PROCEDURE

"Simplex method" is the technical name for the general procedure. Actually there are two slightly different ver-

sions. The original version [10] will really work well only for rather small problems because of the way in which rounding errors build up from step to step. Machine computation of large problems is better carried out by the modified method of Charnes and Lemke.[11]

The general procedure can be worked by hand with the aid of a desk calculator when the number of variables is small, as in the examples discussed in the main text. However, it requires the use of automatic computers in most practical problems owing not to the difficulty but to the sheer quantity of arithmetic involved. Even the simplified avgas problem discussed above required several days of hand computation to solve by the general procedure, while the answer to a problem with twice as many blending stocks and twice as many end products could be obtained in an hour or less on a good electronic computer.

There are still certain limitations on the size of problem which can be handled on existing computers with existing codes of instructions, and some problems which can be solved may cost too much time or money to be worth solving. In many cases, nevertheless, skilled mathematical analysis of a very large problem will show that it can be simplified or broken into manageable parts.

Some problems will undoubtedly remain intractable, but until many more practical applications have been made, it will not really be known whether this will prove to be a frequent obstacle or a very rare one. It should be remembered that rapid progress is being made both in mathematical research [12] and in the design of computers and computing codes. If business finds that it is important to solve problems of linear programing, it seems almost certain that means will be found of solving the great majority of the problems that occur.

[10] See A. Charnes, W. W. Cooper, and A. Henderson, *An Introduction to Linear Programming* (New York, John Wiley & Sons, Inc., 1953).

[11] See *Proceedings of the Association for Computing Machinery* (Pittsburgh, Richard Rimbach Associates, 1952), pp. 97–98.

[12] An important recent advance is to be found in A. Charnes and C. E. Lemke, "Computational Theory of Linear Programming I: The 'Bounded Variables' Problem," O.N.R., Research Memorandum No. 10 (Pittsburgh, Graduate School of Industrial Administration, Carnegie Institute of Technology, 1954).

✦✦✦✦✦✦✦✦✦✦✦✦✦ SOLUTION OF MANAGEMENT PROBLEMS THROUGH MATHEMATICAL PROGRAMMING

The fundamental aim of Operations Research is the introduction of the rigorous methods of the engineer to the analysis of business problems and the process of executive decision making. It will never be possible nor would it be desirable to eliminate completely reliance on executive judgment. But it must be admitted that too many business decisions are based on handy conventions and customs, hunch and guesswork. The tools of Operations Research are designed to help reduce these elements to a minimum. Unfortunately many business problems involve complications which prevent Operations Research analysis by traditional methods and for this reason techniques especially designed to deal with business and related problems are continually worked on by the Operations Researcher. Programming and the particular variety called linear programming are among the new methods which have helped to meet these needs.

Programming techniques have helped the businessman deal with many classes of problems. Programming is essentially a mathematical technique which permits the analyst to determine the best use which can be made of the facilities available to a business firm. Clearly this is precisely the basic problem of optimum decision making and it is to be expected that programming would have wide applicability. It is well to begin by describing some of the major areas where the method has proved its value.

OPTIMUM PRODUCT LINES

When business is satisfactory many firms find they run into a variety of bottlenecks. Factory size, limited equipment capacity, shortage of warehouse space or lack of additional skilled personnel prevent indefinitely large expansion in production. Sometimes it is impossible to eliminate such bottlenecks at least in the short run. Highly trained specialists do not become available overnight especially where the firm does not do the training, and factory expansion requires expansion space, planning time, and construction time. Sometimes the cost of eliminating the bottleneck is not a justifiable expenditure because the expansion of the market is expected to be only temporary. In such circumstances the firm must do the best it can with the facilities which it has available.

The crucial feature of a situation

Cost and Profit Outlook, *May 1956, 9:5, 1–4, with permission of Alderson Associates, Inc., Marketing and Management Counsel.*

such as this is that the production of a relatively unprofitable item takes up valuable and scarce capacity that could have been devoted to the manufacture of items which can bring greater returns. Most firms produce a variety of lines and a variety of sizes and qualities in each line. Which of these the firm should produce and how much of each it ought to make are fundamental policy questions. The businessman is often reluctant to give up or reduce the output of a line for which there is a continuing market and which is able to pay its own way. Yet he must realize that, so long as he is faced with capacity limitations, the real cost to him in making that item includes not only manufacturing costs but also the loss of profits resulting from a decision not to release capacity for the manufacture of other, more remunerative lines.

The solution is not usually to specialize exclusively in the production of the one most profitable item. There are obvious considerations which argue against such a policy. For example, it may be impossible to sell the lucrative lines without offering the customer a more or less full line, and it may therefore even be necessary to provide some items which do not cover their direct costs. The seller must also look ahead to the possibility that general business conditions will someday be less favorable and may therefore decide to keep his way open into less profitable markets which will then be better than no markets at all.

But even neglecting such obvious considerations, concentration on the production of one most profitable item is a poor strategy. Indeed it is not even possible to define what is meant thereby. One item may make good use of machine capacity and will therefore yield the highest profit per scarce machine-hour, whereas another may make more effective use of limited warehouse space. Production of only the former would find warehouses completely loaded before machine time was fully employed, while the latter product, since it is not bulky, might leave warehouses half empty even if the firm's machines were to turn out nothing else.

This problem is clearly of very great importance to business and it is equally clear that there are no easy answers. Certainly guesswork is very unlikely to provide a satisfactory solution. A very considerable part of the work in programming has been successfully applied to solution of problems of this variety.

TRANSPORTATION ROUTING

Another business problem in the solution of which the programmer has been able to help is the selection of transportation routes. Especially where a firm has many plants and its processes involve transshipment of items in various stages of production, substantial savings can be expected from careful planning of commodity movements. If the firm employs its own trucks or other transport facilities, the problem is to route them in such a way that as little mileage as possible is covered in getting the items moved where they have to go.

For example, it will sometimes be possible to cut down on mileage by

shipping from destination *A* to *C* and from *B* to *D* rather than from *A* to *D* and *B* to *C*.

This example involves a possibility of effecting savings which is rather obvious and requires no complex computations. But when there are many shipping points and destinations involved these interrelationships become much too complicated for calculation by instinct and intuition.

Where the firm employs others to do its transporting, the computations may be further complicated by peculiarities in the transportation rate structure, for then the firm's objective is not to minimize ton miles but to minimize payments to the carrier, and the two do not always correspond. Again in this area programming has come to the assistance of the businessman.

MEETING PRODUCT SPECIFICATIONS

Many contracts include a number of minimum specifications which must be met by the product, and sometimes the manufacturer will set up such standards for himself. Usually there is available a variety of ways in which these specifications can be met.

For example, an animal feed may require *X* units of protein per bag, *Y* of carbohydrates and *Z* units of vitamin B, etc. Each of the grains which is put into the animal feed contains some of these nutrients and it is therefore possible to make a bag of feed meet these specifications in many different ways. A very inexpensive ingredient may contain much starch and very little else and to meet the standards it may be necessary to add some more expensive ingredients. But which ingredients should be added and in what proportions? Or will it prove cheaper to begin with somewhat more costly ingredients which supply a better balance of all the nutrients?

The least cost combination of meeting specifications is basically a programming problem. This technique has, for example, been employed in the blending of gasolines.

INVENTORY CONTROL

Operations Research has achieved some of its most noteworthy successes in increasing the efficiency of inventory operations. The results are most likely to be spectacular here because immediate and easily calculable savings can often be expected.

Inventory control procedures usually have several intermediate objectives. They are aimed at determining inventory levels which will even out fluctuations in production, meet unexpected customer demands, and expedite the anticipated flow of products from factory to market in a way which keeps costs down. But here too the rivalry of objectives and various capacity limitations introduces complexity into the problem. A large inventory will keep the probability of disappointed customers and lost sales low but it will clearly also raise warehousing costs. The balancing of these various objectives, the meeting of minimum requirements and the limitations imposed by capacity problems, often calls for programming computations.

OTHER PROBLEMS

Programming techniques have been employed in many other business problems. They can be used to help determine optimum plant and warehouse locations. This is to a large extent a problem of minimizing cost of transportation of raw materials and finished products. For each possible location pattern the minimum transportation cost can be computed by the procedure used to determine optimum routing. The least of these minimum cost figures will determine the best location from the point of view of transport efficiency. Of course other factors such as availability of labor and speed with which markets can be served, must be taken into account in determining where to locate and this, too, adds difficulties.

Programming has also been employed in solving production problems like cutting of paper and cloth in a way which minimizes raw material waste, in the job assignment of specialized personnel, in designing experimental procedures and many other problems.

OPTIMIZATION TECHNIQUES— THE DIFFERENTIAL CALCULUS

Enough has been said to indicate the very great variety of situations to which programming has been applied. There is a common element in all of these which makes them amenable to programming analysis.

To see what is special and novel about programming and what special features characterize the problems to which it can be applied, it must be compared with older methods. Systematic techniques for calculating optimum procedures go back at least to the middle of the seventeenth century and the invention of the differential calculus. In many cases this mathematical method can, given adequate data, indicate precisely what the maximum (profit) or minimum (cost) achievable is.

Sometimes this will be a conceptually straightforward problem like that of pricing a product the sales of which will be decreased by high price. A very low price will then be unprofitable because it will not cover cost, while a very high price will drive the product out of the market. Clearly the most profitable price lies somewhere in between and the calculus will indicate which of the possible in-between prices will be most remunerative. Essentially the trick is to determine a numerical relationship between price and profits and compute the price at which no further price change will add anything to profits. This is optimal because every possible cent of potential profits will then have been squeezed out. Essentially the first derivative which is the cornerstone concept of the calculus may be described as a measure of how much a unit change in the independent variable will add to the value of the dependent variable. In the present case it measures the amount that will be added to profit by a further change in price. The trick in finding the optimum is to set this derivative equal to zero; that is, to find that price at which no more can be added to profit by any further change.

OPTIMIZATION WITH SIDE CONSIDERATIONS

In many problems of optimization there is a further complication in that the outcome, to be acceptable, must meet certain conditions. For example, the problem of fencing in twenty square feet at minimum cost involves the determination of that shape of plot which will save on fencing most effectively. But any saving which is achieved by fencing only nineteen square feet or twenty-one square feet will be unacceptable. This then is essentially a problem of finding the best way of meeting a very precise specification which the mathematician calls "a side condition." So long as the specifications are so precise (the area must be twenty square feet, no more or less, or weight must be just exactly so much, or the starch content of a 100-pound bag of feed must be exactly so many calories, etc.) the optimization problem can still be dealt with by calculus techniques.

However, it is characteristic of many business problems that specifications are not precise but provide only minimum requirements that must be met. Or the specification, rather than stating the precise extent to which a facility will be used, may indicate only the maximum capacity which is available. Any output which overshoots the quality standards or does not fully utilize some part of capacity is not necessarily ruled out. Here the mathematician says the side conditions are inequalities rather than equations. That is, they do not state that X must equal 500 but only that X must be greater than 500.

It is to be noted that this sort of side condition characterizes each of the business problems which has been described earlier in this article. In the optimum product lines and inventory control problems there are maximum capacities to be dealt with. In the problem of meeting specification at minimum cost, each specification is such an inequality. In the transportation routing and plant location problems the presence of such restrictions on the businessman's decisions is less obvious, but they are nevertheless there and play a fundamental role in the computation. There can be capacity limitations on the size and cargo-carrying capacity of the trucks, trains, or ships to be routed. But the more relevant capacity limitation is a peculiar one and only assumes importance in the computational procedure. It states that in no case is it possible to ship less than nothing from one place to another! This rather silly-sounding restriction is important because things like this are never obvious to an electronic computer and, unless it is specifically forbidden to do so, the computer is very likely to assign negative shipments from some supply sources to some destinations. The inequality which states that shipments must be greater than or equal to zero is then necessary to prevent such nonsense computations and this is again an inequality side condition.

WHAT IS PROGRAMMING?

Programming is the mathematical method for the analysis and computation of optimum decisions which do not

violate the limitations imposed by inequality side conditions. For example, it may be most profitable for a firm to produce 10,000 pairs of shoes a week. This is then a sort of optimum. But if the capacity of the firm's plant is only 8,000 pairs of shoes, this "optimum" is an unattainable goal and it becomes necessary to recompute a more modest target which is compatible with the firm's ability to produce. It is thus necessary to compute the optimum among those outputs which are possible; that is, the most profitable of those outputs which do not violate the inequality side condition that production must be less than or equal to 8,000 pairs per week. Programming is the method whereby this information is translated into mathematical form and by which the answer is calculated.

In almost all cases the method of computation is a so-called "iterative procedure." Just as the term "ragout" disguises the fact that it is only stew on a restaurant menu, this fancy term is used to dignify a systematic trial-and-error procedure. The answer to a programming problem will ordinarily not be arrived at directly. Instead the solution is found by groping toward it. But the trial-and-error procedure is not pure guesswork. It is systematic in that it usually involves at least the first two of the following features:

1. There is a mechanical rule which determines, after each step, exactly what the next step is to be on the basis of the results of the trial just completed. The main purpose of this feature of the method of solution is that it makes electronic computation possible. These mechanical brains unfortunately possess no judgment of their own and so they must be told what to do in every contingency or they simply cannot function. This is like teaching a human the rules of algebra before giving him an algebraic problem to solve. In any event, a mechanical rule stating what must be done at each succeeding trial in the trial-and-error procedure is useful because in a complex problem human judgment can go very badly wrong and this can result in a very inefficient, even a totally ineffective, search for the answer.
2. A second feature which usually characterizes the systematic trial-and-error procedure is a guarantee that each trial will yield values which are closer than the last to the correct answer. This very important feature assures the computer that he is always getting closer to his result and is not wasting his time by going off in a totally wrong direction. Of course such a guarantee can only be provided where there is a mechanical rule which specifies step by step what will be done. Otherwise successive steps are unpredictable and it is then not possible to say in advance whether they will be close to or further from the correct answer.
3. For a large class of problems there are available trial-and-error procedure rules which are guaranteed to yield precisely the correct result after a finite number of steps. In other cases where this is not possible, it is possible to calculate a maximum error which says, for ex-

ample, that the result of the most recent trial is at most one tenth of one percent from the correct answer. The desirability of these features of an iterative computational procedure is obvious.

LINEAR PROGRAMMING

It is always possible to use computational methods which are certain to yield a precise answer after a finite number of steps whenever a situation is such that it can be analyzed with the aid of the techniques of *linear* programming. The essential difference between linear and non-linear programming lies in the facts of the situations to which they are applied. For example, in determining which products it is most profitable for the firm to produce, linear programming will apply if a doubling of labor, raw materials, and all other inputs will just permit an approximate doubling of all quantities produced. This will be so where two factories with twice the labor force, etc., can produce twice the output of one factory and production is increased by building more, rather than larger, plants. In other cases economies can be incurred by increasing factory size and here the use of inputs will increase less rapidly than does production. In the latter situation a graph showing the relationship between total quantities of input and total quantities of output will not be a straight line. The mathematical analysis which must be applied here is much more complicated and is called non-linear programming. More generally, linear programming will be applicable to a situation where a given percentage increase in all the independent variables will just suffice to permit the same percentage increase in all the dependent variables.

It is seen that whether or not linear programming will apply to the analysis of a situation cannot be decided arbitrarily by the investigator. Rather it is a matter of the facts of the situation. If these facts happen to be compatible with the relatively simpler description implicit in the models used by linear programming, that method can be used. More often the fact will meet these requirements only roughly, but it may nevertheless be decided to use linear programming because it is believed that the method will yield a sufficiently close approximation to a correct answer.

CAUTIONS TO USERS OF PROGRAMMING

Stubborn refusal to use new techniques and new ideas can be dangerous to the businessman who thereby risks being left behind in the competitive race. It is perhaps equally dangerous to go overboard uncritically on new gimmicks in the name of progress and science. Programming can help the businessman solve some of his problems, but like any other method it can be misleading and dangerous if used without proper planning judgment and careful interpretation of the results.

There is no magic in mathematics—it cannot pull answers out of a hat. Mathematics is only a device which helps deduce the consequences and implications of available information. If that information is incorrect, or the mathe-

matical techniques are not employed properly, the answers will usually be incorrect. In an earlier article in this series ("Selecting an Appropriate Model for an Operations Research Problem," *Cost and Profit Outlook,* Nov. 1955) it was pointed out that mathematical models must usually present an oversimplified version of the facts in order to render the complex situation amenable to analysis. This usually suffices to reduce the answers given by an O.R. computation to the status of approximations. These answers should never be accepted uncritically.

For these reasons no complex methods can be expected to eliminate completely the need for judgment in business decision making. But these methods can make the problems more tractable and reduce the amount of guesswork considerably. An illustration will indicate a way in which this can occur. A linear programming computation may suggest that the output of some commodity should be increased twenty-three per cent. But if the situation is only approximately linear this figure must be taken with a grain of salt. If it is known that the situation is non-linear primarily because there are substantial economies of mass production, it may be inferred that the figure is too conservative and the answer must be revised accordingly. Here, then, an increase in output larger than twenty-three per cent will probably be profitable. In deciding how large an increase is called for the businessman's judgment is therefore indispensable. But this illustration shows how the computation has offered him a guide which can reduce considerably the imponderables of the situation. Where the circumstances are complex, judgment may hardly know where to begin, and the information that an increase in production somewhat larger than twenty-three per cent is called for will be invaluable.

II DYNAMIC PROGRAMMING

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE NATURE AND CHARACTERISTICS OF DYNAMIC PROGRAMMING PROBLEMS *

ABE SHUCHMAN

Dynamic Programming is a relatively new mathematical technique. Originated in 1952 by Richard Bellman, it has been developed rapidly by its creator and his associates at the Rand Corporation. Today, the technique is an important part

* Prepared especially for this volume.

of the tool kit of every operations researcher who is concerned with multiple rather than single decisions.

Executives frequently encounter situations which require them to make a series of decisions with the outcome of each depending on the results of a previous decision in the series. In such situations, it seems apparent that each of the required decisions should take into account its effect on the future as well as its immediate consequences. For example, a production executive would not, as a rule, neglect plant maintenance in order to obtain greater output in one month because he is aware that such an action would reduce output in the following month. In other words, in many problems involving consecutive decisions, the total return resulting from all the decisions may not be optimal at all, if we treat each decision as an independent entity and seek only to achieve the greatest return from it. It may be, instead, that a sacrifice of some gain in making the first decision may make possible far greater gains from the second decision. Dynamic Programming is a technique for determining whether such possibilities exist.

It may be inferred from what has just been said that dynamic programming problems are problems in which a number of interdependent decisions must be made over some span of time. This is certainly true, by and large, for as Richard Bellman himself has said,[1] "we have coined the term 'dynamic programming' to emphasize that these are problems in which time plays an essential role and in which the sequence of decisions is vital." However, as the illustrative articles which immediately follow this one will demonstrate, numerous problems in which time is not a relevant variable can be investigated with the method of dynamic programming. In particular, a decision which involves the allocation of a fixed quantity of resources among a number of alternative uses can be converted into a dynamic programming problem even if there is only the one decision to be made at one moment in time. This can be accomplished by breaking down the decision into several disparate stages so that, in effect, the decision is handled as if it were a sequence of dependent decisions.

The primary characteristic of a dynamic programming problem is, therefore, that it involves a multi-stage process of decision making. Usually the stages are certain time intervals, but not necessarily; they may be stages merely in the technical sense of stages toward solution of the problem.

In addition to the multi-stage process, dynamic programming problems have the following two distinguishing features:

1) At any stage, only a relatively few things must be known in order to describe the problem. For example, a sequence of decisions relating to production would be amenable to analysis by dynamic programming if all one needed to know in order to solve the problem were such things as capacities, inventories at any time and the time remaining to the last decision in the sequence. In other words, dynamic programming problems are characterized by the dependence of the out-

[1] Bellman, Richard, *A General Survey of the Theory of Dynamic Programming*, p. 486, The Rand Corporation, Feb. 11, 1954, p. iii.

come of decisions on a small number of variables.

2) The result of a decision at any stage is merely to alter the numerical values of the small number of variables relevant to the problem. The decision neither increases nor decreases the number of factors on which outcomes depend so that for the next decision in the sequence exactly the same variables must be considered. Thus, to use the production problem as an example again, the effect of a decision will be to alter the amount of capacity, the size of inventories and the time to the last decision, but, as is evident, capacity, inventories, and time remaining continue to be the only critical variables.

Besides a knowledge of the characteristics of dynamic programming problems some terminology is useful. In all situations or problems solved by dynamic programming, there exists a need for making a sequence of decisions and such a sequence of decisions is called a *policy* or a *strategy*. Also, "those policies which are most desirable according to some predetermined criterion will be named *optimal policies.*" [2] Finally, let us add to the features and nomenclature, the basic principle which makes possible the formulation of equations from which optimal policies can be determined. This principle, called by Bellman the *principle of optimality*, is the following: [3]

An optimal policy has the property that whatever the initial state and initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

[2] *Ibid.;* p. vi.

[3] *Loc. cit.*

To assure that the meaning of this principle is clear, let us consider a merchandising firm which buys a single commodity, stocks it in a single warehouse and sells it.[4] The firm's chief executive, we assume, knows both cost and selling price for each of the next five months and is trying to decide how much to buy in each of these five months. Now, our interest is not the determination of an optimal purchase strategy for this executive but rather in elaborating the meaning of the principle of optimality. To this end, suppose the executive makes the wrong decisions for the first and second months, in other words, he buys the wrong amounts of the commodity. Despite this, the method of solution utilized by dynamic programming enables him to begin with the inventory level that exists at the start of the third month and determine an optimal policy for the remaining three months. Wrong decisions in the first and second months do not prevent the making of the right decisions for the third, fourth and fifth months. In short, regardless of what prior decisions have been made, the method of dynamic programming still enables one to find the optimal decisions for the periods or stages that still lie ahead.

The problem just outlined is a highly oversimplified example of one of the many different kinds of business and military problems to which dynamic programming has been successfully applied; namely, procurement problems.

[4] This illustration has been drawn from A. Vazsonyi, *Scientific Programming in Business and Industry*, John Wiley & Sons, Inc., 1958, pp. 227–238, 253–254.

Other problems, for the solution of which the technique has contributed greatly, are equipment replacement policies, production and allocation problems, inventory problems, and control problems. However, all of the many possible uses of this very powerful technique can scarcely be enumerated. Dynamic programming is without question already one of the truly important and useful tools of operations researchers and as more of these researchers master the complexity of the mathematics on which it is based, it will surely find ever wider application.

✦✦✦✦✦✦✦✦✦✦✦✦✦ PLANNING ADVERTISING EXPENDITURES BY DYNAMIC PROGRAMMING METHODS [1]

RICHARD B. MAFFEI

Much of the importance of the new dynamic programming methods has remained unnoticed because of the apparent complexity of the underlying mathematics. This is indeed unfortunate because the essence of the programming methods for certain classes of problems is not new to practicing economists. Dynamic programming methods can provide an efficient computational algorithm for problem types frequently encountered in consumer demand and production theory.

In this paper the following approach will be utilized. First, an illustration concerned with the allocation of an advertising budget among competing media will be presented that will enable the reader to study the relationship between dynamic programming and simple combinatorial analysis. Second, an attempt will be made to show the relationship between demand and production theory on the one hand and programming on the other. Third, an advertising budget allocation problem much like the simpler one discussed in the first section will be cast in terms of the dynamic programming computational procedure and solved. Finally, a very brief mathematical solution will be proffered.

Programming methods of the type to be discussed offer aid and assistance to practicing economists in two ways. First, they offer a means to solve efficiently some very complicated problems. Second, they very frequently offer a conceptual framework that can go a long way toward dispelling misconceptions regarding the essential character of problems.

[1] The research underlying this paper was undertaken as part of a continuing study of "Business Forecasting: Theory and Practice" which is supported by the Sloan Research Fund of the Massachusetts Institute of Technology.

Management Technology, *December, 1960, 1:2, 94–100.*

1. THE COMMON SENSE OF DYNAMIC PROGRAMMING

A manufacturer that prided itself on its scientific approach to management problems decided to try some market experiments in one of its test markets. The company executives wanted to see if they could determine the best combination of media and frequency of appearance of advertisements.

In this particular market there were three media; A, B and C. During any one planning period, an advertisement could appear no more than four times. In medium A, the cost for advertisement was $2,500; in B, $5,000; and in C, $10,000. The total budget allocated to the test was $20,000 for the one experimental period.

Over the past few years, company analysts had collected information about the relationship of sales to frequency of advertisement in the media. They prepared pessimistic and optimistic estimates upon which to base their calculations. These data appear in Table 1.

Since each advertisement in A cost $2,500, the analysts reasoned that at most they could only spend $10,000 in that medium. They could spend the whole budget on four appearances of the advertisement in B at $5,000 per appearance. They could also advertise in C twice and expend the total budget. The question that came up was: What was the combination of frequencies and media that would maximize the expected unit sales?

TABLE 1

ESTIMATED NUMBER OF UNITS OF PRODUCT SOLD PER PERIOD AND FREQUENCY OF ADVERTISEMENTS IN THREE DIFFERENT MEDIA

Frequency per Period	*Estimated Units Sold per Period*					
	Pessimistic			*Optimistic*		
	A	*B*	*C*	*A*	*B*	*C*
1	100	150	250	125	190	300
2	180	260	300	200	290	350
3	240	330	350	255	350	400
4	280	350	400	290	360	450

Each medium served a different market and it could be presumed for example that if a decision were made to advertise twice in B and once in C, total expected sales would be 510 units (i.e. $260 + 250$) under the pessimistic assumption and 590 units (i.e. $290 + 300$) under the optimistic assumptions. With this in mind they prepared the tables of estimated sales based upon pessimistic assumptions (see Table 2).

Suppose for example that a decision was made to advertise twice in A, twice

TABLE 2

TOTAL EXPECTED SALES UNDER VARIOUS COMBINATIONS OF ADVERTISING FREQUENCIES IN MEDIA A, B AND C

Advertisement Frequency of A	*Fc = 0* *Frequency of B Advertisement*					*Fc = 1* *Frequency of B Advertisement*				
	0	*1*	*2*	*3*	*4*	*0*	*1*	*2*	*3*	*4*
0	0	150	260	330	350	250	400	510	580	600
1	100	250	360	430	450	350	500	610	680	700
2	180	330	440	510	530	430	580	690	760	780
3	240	390	500	570	590	490	640	750	820	840
4	280	430	540	610	630	530	680	790	860	880

in B, and once in C. Total expected sales would be 690 (i.e. 180 + 260 + 250). The total cost would be $25,000 (i.e. 2(2,500) + 2(5,000) + 1(10,000)). This combination was not feasible because it violated the budgetary constraints. Table 3 shows the various total costs involved for choosing the various frequency combinations.

Notice that there are a number of total costs that exceed the budget constraint of $20,000 per period. All such combinations, or points, in the table, or "space" are termed "infeasible" in the same sense that they violate basic conditions under which the problem is formulated. Dynamic programming methods are designed to search the "feasible" space efficiently. Let us search the feasible space in greater detail.

There are seven combinations, or points that should be studied more carefully (Table 4). Represent the points by three letters (F_A, F_B, F_C) where F_A, F_B and F_C specify the frequency of advertising in media A, B, and C respectively.

The decision criterion in this case was taken to be the maximization of

TABLE 3

TOTAL COSTS UNDER VARIOUS COMBINATIONS OF ADVERTISING FREQUENCIES IN MEDIA A, B AND C

Frequency of A Advertisement	*Fc = 0* *Frequency of B Advertisement*					*Fc = 1* *Frequency of B Advertisement*				
	0	*1*	*2*	*3*	*4*	*0*	*1*	*2*	*3*	*4*
0	0	5000	10000	15000	20000	10000	15000	20000	25000	30000
1	2500	7500	12500	17500	22500	12500	17500	22500	27500	32500
2	5000	10000	15000	20000	25000	15000	20000	25000	30000	35000
3	7500	12500	17500	22500	27500	17500	22500	27500	32500	37500
4	10000	15000	20000	25000	30000	20000	25000	30000	35000	40000

TABLE 4

TOTAL COSTS AND TOTAL EXPECTED SALES FOR SEVEN KEY COMBINATIONS UNDER BOTH PESSIMISTIC AND OPTIMISTIC ASSUMPTIONS

	Pessimistic Assumptions		*Optimistic Assumptions*	
(F_A, F_B, F_C)	*Total Cost*	*Total Expected Sales*	*Total Cost*	*Total Expected Sales*
(0, 4, 0)	$20,000	350	$20,000	360
(2, 3, 0)	20,000	510	20,000	550
(4, 2, 0)	20,000	540	20,000	580
(0, 2, 1)	20,000	510	20,000	590
(2, 1, 1)	20,000	580	20,000	690
(4, 0, 1)	20,000	530	20,000	640
(0, 0, 2)	20,000	300	20,000	350

total sales. The best combinations under both the pessimistic and optimistic assumptions was (2, 1, 1). That is, two appearances in A, one in B and one in C. This solution maximizes total sales while simultaneously remaining within the budget constraint. Furthermore this solution, as it turned out, was relatively insensitive to differences in basic planning assumptions. Sensitivity analysis can be used to help resolve doubts about inadequacies and errors in the basic data used to solve decision problems.[2] In this illustration even though the data were quite different, the decision reached was the same in each instance.

2. DYNAMIC PROGRAMMING AND ECONOMIC ANALYSIS

The careful reader may have observed some striking similarities between demand analysis utilizing the indifference curve approach, production analysis utilizing the production function and iso-product contour approach, and the methods described in the previous section. In fact they are quite similar, but there is one major dissimilarity.

In both demand and production theory economists generally assume that they have well behaved and continuous functions. These conditions are sometimes specified formally by requiring that first and second derivatives of certain functions exist everywhere. In the illustration in the previous section, a "lumpy," "discontinuous" system was solved in somewhat the same way that the decision rules, marginal utility per dollar equality among products or marginal physical product per dollar equality among factors, are used to resolve problems in demand or production theory. The mathematical economist gets clear rules because he has well-behaved, non-lumpy functions. The practicing economist who is unwilling or unable

[2] A discussion of sensitivity analysis can be found in Richard B. Maffei, "Simulation, Sensitivity and Management Decision Rules," *Journal of Business*, Vol. XXXI, No. 3, July 1958.

to smooth his functions might resort to dynamic programming methods which offer to him nothing really new by way of improved concepts, but something new in terms of computational facility.

It should not be presumed however that it is easy to solve numerical problems of the type discussed. In many complicated problems the use of high-speed computing equipment is required. What dynamic programming provides is a computational procedure that guarantees an optimal solution in a finite number of steps.

3. A PROBLEM IN DYNAMIC PROGRAMMING TERMS

Conside how the problem described earlier could be set up in a slightly different way and under slightly different conditions. Assume that the total budget is \$22,000 rather than \$20,000; that the cost per appearance of A is \$2,000, B is \$6,000 and C is \$10,000. Assume that the estimated units sold relate to the pessimistic assumptions of Table 1 above.

In this illustration as contrasted with the previous one, the total budget need not be expended exactly. For example, two advertisements of C will use up \$20,000 rather than \$22,000. There may be no perfect fit. The budget constraint should be specified as \$22,000 or less. The problem then becomes one to maximize expected total unit sales such that the budget of \$22,000 will not be exceeded.

Table 5 has been set up to span the total budget range of \$22,000 in steps of \$2,000. Consider medium C first (the order is irrelevant and this choice simplifies the calculations). For \$2,000, it would not be possible to purchase

TABLE 5

WORKSHEET FOR AN ADVERTISING ALLOCATION PROBLEM IN TERMS OF DYNAMIC PROGRAMMING

Level of Expenditure	*Medium C*		*Medium B*		*Medium A*	
	Total Sales (1)	*Frequency* (2)	*Total Sales* (3)	*Frequency* (4)	*Total Sales* (5)	*Frequency* (6)
\$2,000	0	***0***	0	0	100	0 *1*
4,000	0	*0*	0	**0**	180	0 1 2
6,000	0	*0*	150	0 *1*	240	0 1 2 3
8,000	0	*0*	150	0 *1*	280	0 1 2 3 *4*
10,000	250	0 *1*	250	0 *1*	280	0 1 2 3 *4*
12,000	250	0 *1*	260	0 1 **2**	280	0 1 2 3 *4*
14,000	250	0 *1*	260	0 1 **2**	430	0 1 2 3 *4*
16,000	250	0 *1*	260	0 1 2	430	0 1 2 3 *4*
18,000	250	0 *1*	330	0 1 2 3	530	0 1 2 3 *4*
20,000	300	0 1 2	330	0 1 2 3	540	0 1 2 3 *4*
22,000	300	0 1 2	330	0 1 2 3	540	0 1 2 3 *4*

an advertisement in C. This holds true for $4,000, $6,000, and $8,000. If the budget were $10,000, one advertisement could be purchased with expected unit sales of 250. For budgets between $10,000 and $18,000, only one advertisement in C could be bought. The italicized numbers under the column headed "Frequency" represent the optimal number of advertisements that could be bought for the stipulated amount.

One of the main features of dynamic programming relates to the nature of the step-by-step optimization procedure. The methods are efficient in the sense that one need only refer, at a given stage of calculations, to the results in the immediately preceding stage. All the relevant information about the system is contained one stage back. This point will be made clearer by referring to columns (3) and (4) of Table 5.

If $10,000 were available, the company could purchase either zero or one advertisement in B. If it bought zero, $10,000 would be left and the previous stage (i.e. columns (1) and (2)) shows that the highest level of sales to be obtained with this amount is 250 units. Notice that it was *not* necessary to specify what combination of advertisements in the various media (here only C obviously) brought the sales about. If the company bought one advertisement in B, expected sales would be 150 units, which is less than 250 and hence not preferable. It is therefore optimal to buy zero units of B when $10,000 is available.

Take the case of $14,000 now. Either 0, 1 or 2 advertisements could be bought in B. If zero were bought, $14,000 would be left giving a best expected level of sales of 250 units (referring to column (1)); if one were bought it would cost $6,000 bringing expected sales of 150 and leaving $8,000 with expected sales resulting therefrom of 0, giving a grand total of 150 which is less than 250; finally if two were bought at a cost of $12,000 bringing expected sales of 260, leaving $2,000 worth zero sales, the grand total would be 260 units which is an improvement over the previous level of 250. Proceeding in this way the whole table can be filled out.

One point must be clarified before the optimal decision is isolated. Medium A cannot be utilized more than four times a period. This is a technical constraint that can be easily introduced into the system. It is apparent that if the data were available it would be possible to study eleven or fewer appearances of advertising in Medium A. This would introduce computational rather than conceptual difficulty.

Studying the last column opposite the figure $22,000 in Table 5 we see that the number four is italicized. Four appearances of an advertisement in A would cost $8,000. The optimal use of the residual $14,000 can then be ascertained by going to the column (4) entry opposite $14,000. This figure shows that two advertisements in B should be used. Those advertisements would cost $12,000, giving a total expenditure thus far of $8,000 and $12,000, or $20,000. The residual $2,000 can be seen to permit no advertisements in A.

The best strategy here is to use four ads in A, and two in B. The total cost would be \$2,000 less than budget. The total expected sales resulting therefrom would be 540 units (i.e. 280 + 260).

4. THE MATHEMATICS OF DYNAMIC PROGRAMMING

NOTATION:

let

1. d : represent an assumed level of expenditure such that $0 \leqq d \leqq B$.
2. L : represent the budget limit.
3. K_i : represent the cost per advertisement in medium i where $i = A, B, C$.
4. $h_s(d)$: represent the optimal expected sales at stage s, and level of expenditure d.
5. F_i : represent the number of times during the period that an advertisement appears in medium i.
6. $G_i(F_i)$: sales as a function of the frequency of advertising in medium dium i.

STRUCTURE OF THE DECISION PROBLEM

The decision problem is to find the combination (F_A^*, F_B^*, F_C^*) that maximizes total expected sales and does not violate the budgetary constraint.

COMPUTATIONAL PROCEDURES

The step by step procedure utilized in the illustration was based on the following formulae:

(a) $h_1(d) = \max g_c(F_C)$ for all d

where $0 \leqq F_C \leqq \left[\frac{d}{K_C}\right]$

and $\left[\frac{d}{K_C}\right]$ is an integer found by ignoring any decimal part of the number.

(b) $h_2(d) = \max \{G_B(F_B) + h_i(d - K_B F_B)\}$

where $0 \leqq F_B \leqq \left[\frac{d}{K_B}\right]$

(c) $h_3(d) = \max \{G_A(F_A) + h_2(d - K_A F_A)\}$

where $0 \leqq F_A \leqq \left[\frac{d}{K_A}\right]$

After the table was filled in, it was simple to find the value F_A^*. Call this value L_A. [The budget remaining after the decision was made for A.] Then going into stage 2 from stage 3 with L_A equal to the appropriate d, it was possible to obtain F_B^*. The residual remaining was then $L_A - K_B F_B^* = L_B$. By a similar set of steps F_C^* was obtained.

The remaining residual was found by the algebraic relation

$$R = L - K_A F_A^* - K_B F_B^* - K_C F_C^*$$

The total expected sales were obtained directly by observing the value opposite L in the last column and the calculations were checked by finding

$$S = G_A(F_A^*) + G_B(F_B^*) + G_C(F_C^*)$$

This completes the mathematical presentation.

5. THE FLEXIBILITY OF DYNAMIC PROGRAMMING

The illustration just presented is a rather simple one and it would be unfortunate indeed if the reader concluded that such was the character of the programming techniques.

There is here only a suggestion of the possible use of these methods. It is feasible to expand this illustration to include an arbitrarily large number of media with an equivalent number of technical constraints referred to earlier. It is possible to impose various other constraints upon the system and still have a fair amount of computational ease.

Many kinds of problems can be framed in dynamic programming terms. Problems involving various kinds of promotional forms other than advertising such as sales force size, special promotions and so on can be framed and solved. It is also possible to study the timing of advertising activity. Only one of many possible uses has been indicated here.

✦✦✦✦✦✦✦✦✦✦✦✦✦ DYNAMIC PROGRAMMING

ANDREW VAZSONYI

1. DISTRIBUTION OF SALES EFFORT BETWEEN VARIOUS MARKETING AREAS

. . . In order to simplify matters let us start with a simple problem, where our firm has only two marketing areas to consider and the problem is to distribute a given number of salesmen between these two marketing areas. The profit for each of these marketing areas is given as a function of the sales effort expended, as shown in Fig. 1 and Fig. 2. For instance, it can be seen in Fig. 1 that if 7 salesmen are assigned to the first marketing area a profit of $96,000 results. Figure 1 also shows that, if more than 8 salesmen are assigned to this marketing area, then profits will actually go down. This is so because if more than 8 salesmen work in the area they do not get more sales, but they all have to be paid, and so the net profit for the corporation will go down. As a matter of fact, the situation could conceivably arise where too many salesmen antagonize the customers and sales even drop.

The problem here is to distribute a given number of salesmen between the two marketing areas so that the profit

Reprinted with permission from Andrew Vazsonyi, Scientific Programming in Business and Industry, *1958, 219–227, John Wiley and Sons, Inc.*

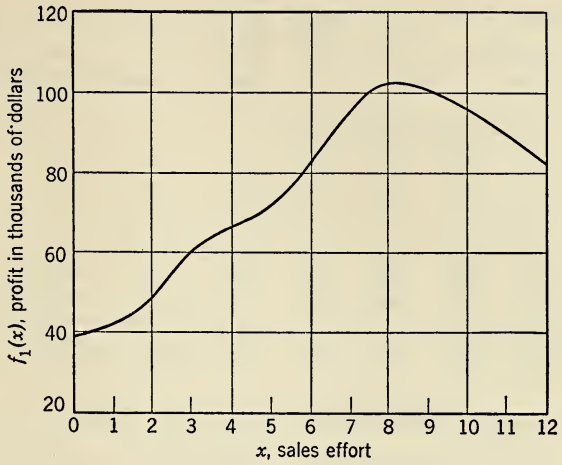


FIGURE 1

PROFIT FOR FIRST MARKETING AREA AS FUNCTION OF SALES EFFORT

will become maximum. We will use $f_1(x)$ to denote profit in the first marketing area if x number of salesmen are employed in this area, and $f_2(x)$ to denote profit in the second marketing area if x number of salesmen are employed there.

As an illustration, let us suppose that the corporation has 6 salesmen, and the problem is to allocate these 6 salesmen so that profit will be maximum. There are only seven possibilities: we can allocate no salesmen to area 1 and 6 salesmen to area 2; we can allocate

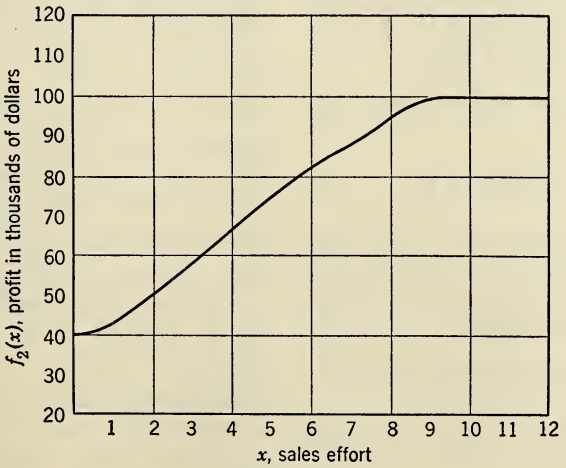


FIGURE 2

PROFIT FOR SECOND MARKETING AREA AS FUNCTION OF SALES EFFORT

1 salesman to area 1 and 5 salesmen to area 2; etc. Therefore, we can prepare the table shown.

x_1	0	1	2	3	4	5	6	
x_2	6	5	4	3	2	1	0	
$f(x_1)$	38	41	48	58	66	72	83	In thousands of dollars
$f(x_2)$	82	75	66	60	50	42	40	
z	120	116	114	118	116	114	123	

The last row shows the profit realized corresponding to each of the seven allocation schemes. It can be seen that the best allocation is to assign all of the 6 salesmen to the first area and no salesmen to the second area, because this results in maximum profit of $123,000. To use our mathematical notation, we can say that the method of obtaining the best allocation is the selection of the largest of the following seven numbers:

$[f_1(0) + f_2(6)]$, $[f_1(1) + f_2(5)]$, $[f_1(2) + f_2(4)]$, $[f_1(3) + f_2(3)]$, $[f_1(4) + f_2(2)]$, $[f_1(5) + f_2(1)]$, $[f_1(6) + f_2(0)]$

A simpler way to express this maximum profit z is

$$z = \max_{0 \leq x \leq 6} [f_1(x) + f_2(6 - x)] \qquad (1)$$

Let us now denote by $F(A)$ the maximum profit that can be realized if A salesmen are allocated (the optimum fashion) between the two marketing areas. Then we can write equation (1) as

$$F(6) = \max_{0 \leq x \leq 6} [f_1(x) + f_2(6 - x)] \qquad (2)$$

We can proceed now to determine the maximum profit for 1, 2, 3, $\cdots$ 12 salesmen; in fact we can determine the profit $F(A)$ for any number of salesmen. A somewhat simpler procedure is shown in Table 1, where the profit is computed by assuming that a certain number of salesmen are assigned to the first area, and that a certain number are assigned to the second area. For instance, it can be seen that, if 3 salesmen are assigned to the first area and 2 salesmen to the second area, then the profit realized will be $110,000. Now observe the diagonals (Table 1), as along these diagonals the combined number of salesmen assigned to the two areas is the same. For instance, if we assign 4 salesmen between the two areas we read along the diagonal the following numbers: 104, 99, 98, 102, 106. We see, then, that the best way to allocate these 4 salesmen is to allocate them all to the first marketing area; the profit in this case is $106,000. Therefore, by using a table of this kind we can determine $F(A)$, which is the maximum profit realizable if a combined number of A salesmen are assigned to the two marketing areas. Mathematically speaking, we have described here a method of determining $F(A)$ from the equation

$$F(A) = \max_{0 \leq x \leq A} [f_1(x) + f_2(A - x)] \qquad (3)$$

TABLE 1

OPTIMUM DISTRIBUTION OF SALES EFFORT BETWEEN TWO MARKETING AREAS

Number of Salesmen in Second Area		Number of Salesmen in First Area												
		0	1	2	3	4	5	6	7	8	9	10	11	12
		38	41	48	60	66	72	83	96	102	100	95	89	82
0	40	78*	81*	88*	100*	106*	112	123*	136*	142*	140	135	129	122
1	42	80	83	90	102	108	114	125	138	144	142	137	131	
2	50	88	91	98	110	116	122	133	146*	152	150	145		
3	58	96	99	106	118	124	130	141	154*	160	158			
4	66	104	107	114	126	132	138	149	162*	168				
5	75	113*	116	123	135	141	147	158	171*					
6	82	120	123	130	144	150	154	165						
7	88	126	129	136	138	154	160							
8	95	133	136	143	155	161								
9	99	137	140	147	159									
10	100	138	141	148										
11	100	138	141											
12	100	138												

Profit in thousands of dollars

(The numbers marked with asterisks are the maxima along each diagonal.)

$F(A)$ can be represented either in a tabular form or can be plotted as shown in Fig. 3.

Thus, as long as we deal with two marketing areas, we can solve the distribution of effort problem, irrespective of the shape of profit functions. Let us suppose, now that there are three marketing areas to consider. The profit as a function of sales effort for the third marketing area is shown in Fig. 4. How should we solve the problem in this

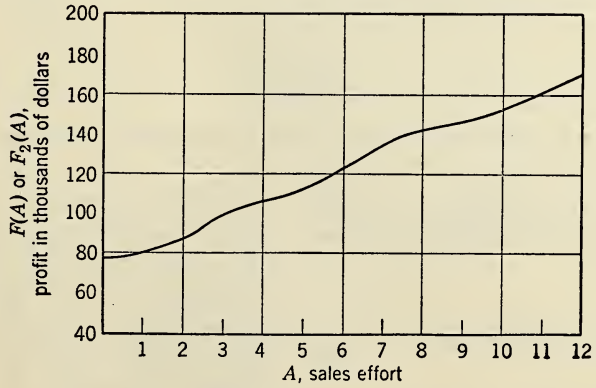


FIGURE 3

PROFIT FOR OPTIMUM ALLOCATION OF SALES EFFORT BETWEEN TWO MARKETING AREAS

more complicated case? A simple trial and error computation shows that 6 salesmen can be allocated 21 different ways and 12 salesmen can be allocated 78 ways. As we are trying to develop a general method of solving the problem, so that not only three or four but any number of marketing areas can be considered, it is obvious that the method of preparing tables for all possible cases leads to an enormous amount of computation.

problem, we propose to consider the situation from a somewhat different point of view. Suppose we want to solve the problem of allocating 6 salesmen to three marketing areas, with the proviso *that we are allocating 2 salesmen to the first two areas and the rest (that is, the 4 salesmen that remain) to the third marketing area.* What is the best possible profit that can be realized in this special problem?

The best profit that can be obtained

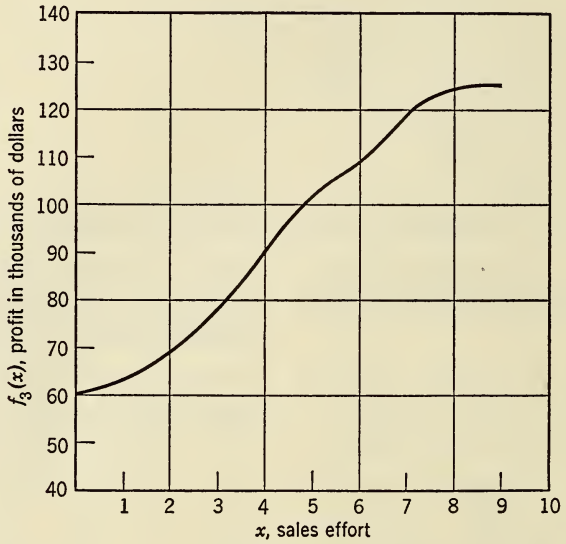


FIGURE 4

PROFIT FOR THIRD MARKETING AREA AS FUNCTION OF SALES EFFORT

Mathematically speaking, the problem is to obtain the maximum profit as represented by

$$z = \max [f_1(x_1) + f_2(x_2) + f_3(x_3) + \cdots f_N(x_N)] \qquad (4)$$

where N, the number of terms on the right-hand side, relates to the number of marketing areas to be considered.

In order to obtain a solution to this

by using 2 salesmen in the first and second marketing area is given by $F(2)$, and the profit from allocating 4 salesmen to the third area is, of course, $f_3(4)$. Therefore, we can say that the best profit for this particular problem is given by

$$z = F(2) + f_3(4) \qquad (5)$$

However, there is no good reason to allocate exactly 2 salesmen to the first

two areas, and so we abandon this assumption. We should try to allocate 0 salesman, 1 salesman, 2 salesmen, etc., to the first two marketing areas. Therefore it can be seen that our problem is to select the largest of the following seven numbers:

$$[F(0)+f_3(6)],\quad [F(1)+f_3(5)],\quad [F(2)+f_3(4)],\ [F(3)+f_3(3)],\ [F(4)+f_3(2)],\ [F(5)+f_3(1)],\ [F(6)+f_3(0)] \qquad (6)$$

Mathematically speaking, we see that we can solve the problem of the three marketing areas by evaluating the right-hand side of the following equation:

$$z = \max_{0\leq x\leq A} [F(x)+f_3(A-x)] \qquad (7)$$

Now we propose to introduce a new notation. Let us denote by $F_2(A)$ the best profit that can be realized by allocating A number of salesmen between the first two marketing areas. (The subscript 2 reminds us that we are dealing with *two* marketing areas.) This is the same function that we denoted by $F(A)$ before, and we recall that, according to equation (3), this function is given by the following equation:

$$F_2(A) = \max_{0\leq x\leq A} [f_1(x)+f_2(A-x)] \qquad (8)$$

Let us now denote by $F_3(A)$ the best profit that can be realized by allocating A number of salesmen between the three marketing areas. With this notation, then, equation (7) changes into the following:

$$F_3(A) = \max_{0\leq x\leq A} [F_2(A)+f_3(A-x)] \qquad (9)$$

It can be seen, then, that the problem with three marketing areas can be solved in exactly the same fashion as the problem for the two marketing areas. We present Table 2 to show the

TABLE 2

OPTIMUM DISTRIBUTION OF SALES EFFORT FOR THREE MARKETING AREAS

Number of Salesmen in Third Area		Combined Number of Salesmen in First and Second Area												
		0	1	2	3	4	5	6	7	8	9	10	11	12
		78	81	88	100	106	113	123	136	142	146	154	162	171
0	60	138	141	144	160*	166	173	183	196	202*	206	214	222	231
1	64	142*	145	148	164	170	177	187	200	206	210	218	226	
2	68	146*	149	152	168	174	181	191	204	210	214	222		
3	78	156	159	162	178	184	191	201	214	220	224			
4	90	168*	171	174	190	196	203	213	226*	232				
5	102	180*	183	186	202*	208	215*	225	238*					
6	109	187*	190	193	209*	215*	222	232						
7	119	197*	200	203	219	225	232							
8	124	202*	205	208	224	230								
9	125	203	206	209	225									
10	125	203	206	209										
11	125	203	206											
12	125	203												

Profit in thousands of dollars

(The numbers marked with asterisks are the maxima along each diagonal.)

method of solution. The top row shows the combined number of salesmen that are allocated to the first and second area; the second row shows the maximum profit that can be realized by allocating the salesmen between the first two areas in the best possible fashion. (In other words, the second row is $F_2(A)$.) The first column on the left-hand side shows the number of salesmen allocated to the third marketing area, and the next column shows the profit that can be realized by allocating these salesmen to the third area. The table itself shows the combined profit. What we have to do again is to follow the diagonals and select the maximum value. In Table 2 these maxima are shown by asterisks. Figure 5 shows a graphical representation of $F_3(A)$, or the best profit that can be realized when considering three marketing areas.

The problem of four marketing areas represents nothing new as we can simply write

$$F_4(A) = \max_{0 \leq x \leq A} [F_3(A) + f_4(A - x)] \quad (10)$$

This problem can be solved again by preparing an appropriate table.

Mathematically speaking, the problem is to solve the following equation:

$$z = \max_{x_1 + x_2 + \cdots + x_N = A} [f_1(x_1) + f_2(x_2) + \ldots + f_N(x_N)] \quad (11)$$

which is the same as (4) but the side condition is stated explicitly here. The method of solution is to obtain $F_2(A)$ from equation (8), $F_3(A)$ from equation (9), $F_4(A)$ from equation (10), etc. In general, for any number of marketing areas, we obtain the solution from

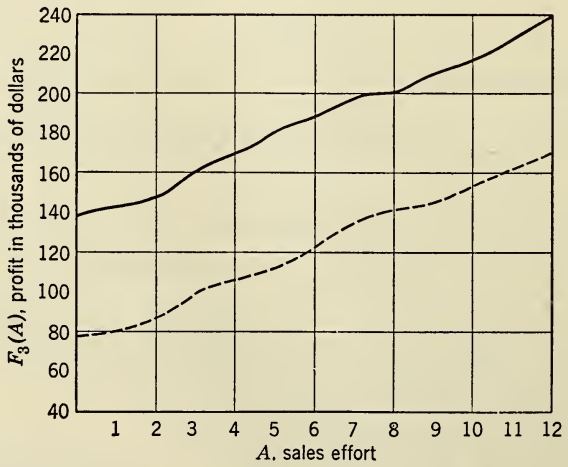


FIGURE 5

PROFIT FOR OPTIMUM ALLOCATION OF SALES EFFORT FOR THREE MARKETING AREAS

(The dotted line shows the maximum profit when the salesmen are allocated to the first two areas.)

$$F_n(A) = \max_{0 \leq x \leq A} [F_{n-1}(A) + f_n(A - x)]$$

$$n = 2,\ 3, \cdots \qquad (12)$$

In the mathematical literature such a method of solution is called an *algorithm*. Solution with the aid of an algorithm means that the solution is not obtained at once, but is obtained step by step, and in each step the results obtained in the preceding steps are utilized. The first step is to obtain the solution when there are two marketing areas; this solution can be obtained with the aid of a table. Then step by step more tables are built up and the solutions successively obtained.

III SYMBOLIC LOGIC

✦✦✦✦✦✦✦✦✦✦✦✦✦ SYMBOLIC LOGIC: *a summary of the subject and its application to industrial engineering*

SANFORD S. ACKERMAN

A major aim of anyone engaged in a pure or applied science is to state propositions clearly and describe systems without ambiguity or the possibility of misinterpretation. Ordinary language is at best a difficult medium through which complicated propositions may be concisely expressed, and it fails miserably when relationships are mathematically or logically complex. Early Egyptian mathematicians struggled with the problem of what number plus one-fifth of itself equals twenty-one for several hundred years (8). They had no symbols for digits or algebraic operators and, therefore, had to state the problem and attempt its solution using only words. The modern mathematician uses a formal system of digits, symbols, and operator notation to free himself of the encumbrances of words. It is apparent that some method of expression is needed to free the scientist from the tyranny of words and permit him to express propositions and describe complex systems in a rigorous manner, analogous to the way he can express mathematical ideas.

Symbolic logic is a means of expressing verbal propositions and statements of relationships between propositions in a concise and unambiguous form. The manipulation of propositions stated

Reprinted from the September–October 1957 issue of The Journal of Industrial Engineering, *8:5, 293–299, Official Publication of the American Institute of Industrial Engineers, Inc., 345 East Forty-Seventh Street, New York 17, New York.*

in symbolic form is performed by methods of Boolean algebra. It is the object of this paper to present the concept of the calculus of propositions—the exposition of verbal statements and their relationships in symbolic form—and Boolean algebra—the algebra of classes.

The uses of logic, like the uses of mathematics, cannot be limited to specific applications in a particular field. Just as the same differential equation can describe the mechanical vibrations of a physical system or the flow of electricity in a circuit, a symbolic sentence can describe the conditions of a contract or the operation of a computer element. Some of the examples of applications will be presented to show the manner in which symbolic logic can be used as an analytical method, with the inference that a substitution of some element of the industrial scene for the subject of the example can make the technique applicable to particular situations in . . . [business.]

In 1854, George Boole, an English mathematician, wrote the treatise, "An Investigation of The Laws of Thought, on which are Founded the Mathematical Theories of Logic and Probabilities" (3). He proposed a system by which propositions could be represented by symbols and manipulated to draw logical conclusions. Boole's original symbolism has been modified and extended since that original publication. In 1913, Alfred North Whitehead and Bertrand Russell, using the somewhat more complicated notation of Giuseppe Peano, an Italian mathematician, wrote the monumental work, "Principia Mathematica," in which they attempted to explore the logical foundations of mathematics, asserting that mathematics was in reality a branch of the more primary discipline, logic (8). For the most part, however, formal symbolic logic remained an academic pursuit without application in the engineering or business worlds. The first reported non-academic application of symbolic logic was by Edward C. Berkely at the Prudential Life Insurance Company in 1936 (8). Using symbolic logic, he examined the profusion of complicated rules governing the payment of premiums, and discovered that there were conflicts in the rules. An investigation of the files revealed that there were cases which had been handled under the conflicting rules. Since then, symbolic logic has been used to find ambiguities in contracts, and has found extensive application in the design of computer circuits (7).

SYMBOLIC LOGIC AND BOOLEAN ALGEBRA

Before proceeding further, it is desirable that the reader acquaint himself with the basic laws and definitions of Boolean algebra, and with the use of logical symbolism. Operator notation is not standardized; therefore, as each symbol is defined in the manner it is used in this paper, alternate forms are shown alongside in parentheses.

A set, or class, represents all elements lying within the specified bound which defines the set. Boolean algebra is an algebra of sets or classes of elements which have the properties described below (2):

1. Two binary operations, cup and cap, $\cup$ — Cup $(+, \vee)$; $A \cup B$ reads, "A or B, or both," (Union)
$\cap$ — Cap $(\times, \wedge)$; $A \cap B$ reads, "both A and B," (Intersection) which satisfy the following laws:

   Indempotent: $A \cap A = A$, and $A \cup A = A$ (from which follows, $A^n = A$)

   Commutative: $A \cap B = B \cap A$, and $A \cup B = B \cup A$

   Associative: $A \cap (B \cap C) = (A \cap B) \cap C$, and $A \cup (B \cup C) = (A \cup B) \cup C$

   Distributive: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

2. Two elements 0 and 1, which are universal bounds and satisfy the laws of union and intersection; i.e., $0 \leqq A \leqq 1$ for all A (Universal Bounds).
$0 \leqq A \leqq 1$ reads, "the null set is contained in A is contained in the universal set."

   $0 \cap A = 0$, and $1 \cap A = A$ (Intersection)

   $0 \cup A = A$, and $1 \cup A = 1$ (Union)

   These relationships are more easily understood with the aid of Venn's Diagram:

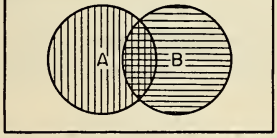


FIGURE 1

Everything within the bound = the universal set = 1. $A \cup B$ (read, "union of A and B") = everything that is either A or B, or both; i.e., the areas covered by vertical lines, horizontal lines, and cross-hatching. $A \cap B$ (read, "intersection of A and B") = the area covered by cross-hatching. 0 = the null set; i.e., that set which contains nothing.

3. A unary operation of complementation which obeys the complementary, dualization, and involution laws;

$$A \cap A' = 0, \text{ and } A \cup A' = 1 \quad \text{(Complementarity)}$$
$$(A \cup B)' = A' \cap B', \text{ and } (A \cap B)' = A' \cup B' \quad \text{(Dualization)}$$
$$(A')' = A \quad \text{(Involution)}$$

A' reads, "not A." $'$ and $\sim$ are the symbols of negation. The dualization law may be stated in general as: The "not" of a polynomial may be found by taking the "not" of each term, and then interchanging cup and cap operators.

The following are some examples of manipulation of logical expressions (7):

(a) $A \cup (A \cap B) = A \cap (1 \cup B)$
$= A \cap 1$
$= A$

(b) $A \cap (1 \cup B) = (A \cap A) \cup (A \cap B)$
$= A \cup (A \cap B)$
$= A$ (from a)

(c) $(A' \cap B') \cap (A \cap B)' = (A' \cap B') \cap (A' \cup B')$
$= (A' \cap A' \cap B') \cup (A' \cap B' \cap B')$
$= (A' \cap B') \cup (A' \cap B')$
$= A' \cap B'$

$$
\begin{aligned}
\text{(d)}\quad A \cup (A' \cap B) &= [(A \cup A' \cap B')']' \\
&= [A' \cap (A \cup B)]' \\
&= [A' \cap A \cup A' \cap B]' \\
&= [0 \cup (A' \cap B')]' \\
&= A \cup B
\end{aligned}
$$

$$
\begin{aligned}
\text{(e)}\quad (A \cap B) \cup (B' \cap C') \cup (A \cap C') &= (A \cap B) \cup (B' \cap C') \cup (A \cap C') \cap (B \cup B') \\
&= (A \cap B) \cup (B' \cap C') \cup (A \cap B \cap C') \cup (A \cap B' \cap C') \\
&= [(A \cap B) \cup (A \cap B \cap C')] \cup [(B' \cap C') \cup (B' \cap C' \cap A)] \\
&= [(A \cap B) \cap (1 \cup C')] \cup [(B' \cap C') \cap (1 \cup A)] \\
&= (A \cap B) \cup (B' \cap C')
\end{aligned}
$$

The above manipulations may be more easily followed when the cup and cap operators are thought of as the addition $(+)$ and multiplication $(\times)$ signs, respectively, of ordinary algebra, for which cup and cap are somewhat analogous.

The translation of ordinary language to symbolic form requires the interpretation of the above operators, and the definition of two more symbols. In a verbal context, "$A \cup B$" denotes the statement, "that which is either A, or B, or both." "$A \cap B$" denotes the statement, "that which is both A and B." "A'" denotes the statement, "not A," or, "it is not true that A. . . ." The first type of statement is disjunction, the second conjunction, and the third, negation. Another form of statement is, "$A \rightarrow B$" (or "$A \supset B$"); i.e., "if A, then B." This form is described as conditional, or a statement of material implication. Equivalence is shown as, "$A \equiv B$" (or "$A \longleftrightarrow B$"), which reads, "A if, and only if B," or, "A is equivalent to B." [1] Material implication may also be expressed in terms of other operators for purposes of manipulation: $A \rightarrow B = A' \cup B$.

It is sometimes of interest to determine the truth or falsity of an entire statement when the constituent elements take on all possible combinations of truth and falsity. A convenient graphical representation is the truth table. Shown below are truth tables for conjunction, disjunction, material implication, and equivalence.

A	B	A ∩ B	A ∪ B	A → B	A ≡ B
T	T	T	T	T	T
T	F	F	T	F	F
F	T	F	T	T	F
F	F	F	F	T	T

For instance, given the statement $(A \cap B) \rightarrow (C \cup D)$, and the values A true, B, C, and D false, find the truth value of the entire statement. For the stated values in the table above, $(A \cap B)$ is false and $(C \cup D)$ is false. The given statement may now be thought of as an equivalent statement $E \rightarrow F$ where E and F are both false. For these values in the table, it is seen that the value of the whole expression is true.

[1] The reader should note that some of these interpretations are only for ease of conception. A more rigorous interpretation is required when complex relationships are treated. See (9), page 93.

APPLICATIONS

. . . Symbolic logic may also be applied to the resolution of complicated verbal relationships; e.g., insurance policy and contract stipulations. Two examples of this type of application follow. The first is from a book on logic by Lewis Carroll (8).

No kitten that loves fish is unteachable. No kitten without a tail will play with a gorilla. Kittens with whiskers always love fish. No teachable kitten has green eyes. No kittens have tails unless they have whiskers.

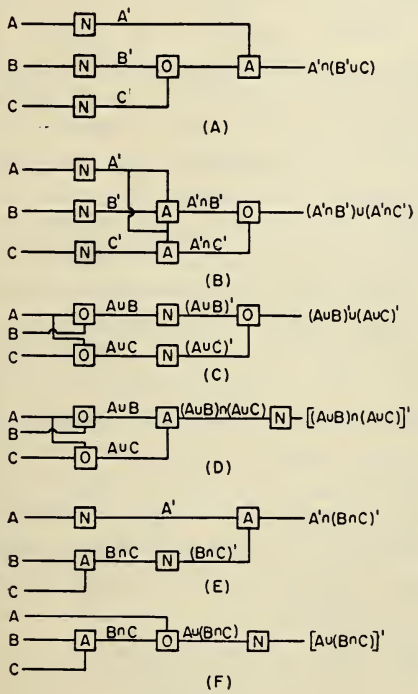


FIGURE 2

Will green-eyed kittens play with a gorilla? If we agree that the complement of each of the foregoing statements is also true, we may resolve the problem as follows: The elements are first defined by symbols: Let kittens which:

love fish = b
are teachable = c
have tails = d
will play with a gorilla = e
have whiskers = f
have green eyes = g,

and the verbal conditions written as symbolic sentences [2]

1. $b \rightarrow c$ or $c' \rightarrow b'$
2. $e \rightarrow d$ or $d' \rightarrow e'$
3. $f \rightarrow b$ or $b' \rightarrow f'$
4. $g \rightarrow c'$
5. $d \rightarrow f$ or $f' \rightarrow d'$

Question: $g \rightarrow e$?

$$g \rightarrow c' \rightarrow b' \rightarrow f' \rightarrow d' \rightarrow e' \therefore g \rightarrow e'$$
$$4. \quad 1. \quad 3. \quad 5. \quad 2.$$

which says, "if a kitten has green eyes, then he will not play with a gorilla."

The previous example may be solved by carefully rewriting and rearranging the original relationships in verbal form. However, relationships are often confusing and complex to a degree which makes solution impossible by methods of verbal reasoning. Consider the following:

If a mathematician does not have to wait 20 minutes for a bus, then he either likes Mozart in the morning or whiskey at night, but not both. If a man likes whiskey at night, then he either likes Mozart in the morning and does not have to wait 20 minutes for a bus, or he does not like Mozart in the morning and has to wait 20 minutes

[2] The usual convention of logicians to use capital and lower case letters to denote statements and sets, respectively, is not adhered to in this paper.

for a bus or else he is no mathematician. If a man likes Mozart in the morning and does not have to wait 20 minutes for a bus, then he likes whiskey at night. If a mathematician likes Mozart in the morning, he either likes whiskey at night or has to wait 20 minutes for a bus; conversely, if he likes whiskey at night and has to wait 20 minutes for a bus he is a mathematician—if he likes Mozart in the morning. When does a mathematician wait 20 minutes for a bus (8)?

Proceeding as in the previous problem: A man who:

is a mathematician = A
likes whiskey at night = B
likes Mozart in the morning = C
waits 20 minutes for a bus = D

(a) $A \cap D' \rightarrow (B \cup C) \cap (B \cap C)'$
(b) $B \rightarrow (C \cap D') \cup (C' \cap D) \cup A'$
(c) $C \cap D' \rightarrow B$
(d) $A \cap C \rightarrow (B \cup D)$
(e) $B \cap D \rightarrow (C \rightarrow A)$

When a mathematician waits 20 minutes for a bus, it is equivalent to saying that A and D are true. If a truth table is constructed to show how the truth values of the problem statements vary for all possible combinations of truth values of B and C, the solution of the problem may be found by choosing those values of B and C for which all the problem statements are true; i.e., the conditions of A, B, C, and D which define the situation in which a mathematician waits 20 minutes for a bus.

A	B	C	D	a	b	c	d	e
T	T	T	T	T	F			
T	F	T	T	T	T	T	T	T
T	T	F	T	T	T	T	T	T
T	F	F	T	T	T	T	T	T

The three solutions say, $A \cap D \rightarrow (B' \cap C) \cup (B' \cap C') \cup (B \cap C')$ which reduces to, $A \cap D \rightarrow B' \cup C'$. The verbal statement of the solution is, "a mathematician waits 20 minutes for a bus when he does not like whiskey at night, or when he does not like Mozart in the morning, or both."

The possible complexity of the language problem is evident from the relative difficulties of the . . . examples. The extreme difficulty of finding conflicts in lengthy and complex legal documents, or in industrial specifications by verbal methods indicates that symbolic logic is a tool worthy of consideration for application in this area.

Boolean algebra is an algebra of classes, and, therefore, a useful method for solution of problems involving numerical classifications. Consider the following example from the joint associateship examination for actuaries, 1935: Certain data obtained from a study of a group of 1,000 employees in a cotton mill as to their race, sex, and marital status were unofficially reported as follows: 525 colored lives; 312 male lives; 470 married lives; 42 colored males: 147 married colored; 86 married males; 25 married colored males. Are the data consistent? Let $N(A)$ denote the number of elements in class A. From the definition of union and intersection, the operator N has the property

$$N(A \cup B) = N(A) + N(B) - N(A \cap B).$$

Let:

c = colored lives
m = male lives
w = married lives

Using the property of the operator twice,

$$\begin{aligned} N(c \cup m \cup w) &= N(c) + N(c \cup w) \\ &\quad - N[(c \cap m) \cup \\ &\quad (c \cap w)] \\ &= N(c) + N(m) \\ &\quad + N(w) - N \\ &\quad (m \cap w) \\ &\quad - N(c \cap m) \\ &\quad - N(c \cap w) + N \\ &\quad (c \cap m \cap w) \end{aligned}$$

Substituting from the above data,

$$\begin{aligned} &= 525 + 312 + 470 - \\ &\quad 42 - 147 - 86 + 25 \\ &= 1{,}057 \end{aligned}$$

For the conditions to be fulfilled, there would have to be 1,057 employees. Since there are only 1,000 employees, the data are inconsistent (2).

The above examples of applications of symbolic logic have shown how relatively simple verbal relationships which would be extremely difficult to manipulate and simplify by using ordinary language, can be handled easily when symbolic logic is used. However, when problems are considerably more complex, and the number of interrelationships grow, the task of manipulating logical equations and deriving truth tables reaches formidable proportions. Mechanical and electrical devices have been built to test symbolic logic sentences (5). One class of these could be called logical analog computers. For instance, a simple analog for the sentence $A \rightarrow (B \cap C) \cup D$ could be constructed from relays as shown in Figure 3.

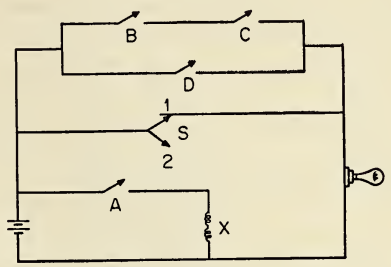


FIGURE 3

The switch S is excited by coil *X;* it is in position 1 except when *X* is energized by *A* being closed. *A, B, C, D,* and *E* are relays whose coils are energized when the value of the corresponding term is "true." E.g., take the case of *A* false, *B* true, *C* false, and *D* false. Relay *B* is closed; *A, C,* and *D* are open; switch S is in position 1; and the bulb lights—indicating that the sentence is true for the given truth values of the constituents. An entire truth table for this logical sentence can be generated by systematically trying all the combinations of constituent truth values. General purpose logical analog computers could be constructed of banks of relays, or electronic switching circuits. The design of these analogs is the inverse of the logical design of computer elements mentioned earlier.

A method of generating truth tables from logical sentences, using a digital computer, is reported in (1). Programs which generate truth values for logical sentences of two variables in any of the forms—conjunction, disjunction, material implication, equivalence, and negation—are used on "nested" logical sentences to generate truth tables of sentences of any length or complexity. This nesting technique is shown as follows (1):

$$((p_1 \rightarrow p_0) \cap (((p_0 \cap p_1) \cap p'_3) \rightarrow p_2)) \cap ((p_2 \rightarrow p_1) \cap (p'_0 \rightarrow p_3))$$

$a_1 \equiv (p_1 \rightarrow p_0)$ $\quad b_1 \equiv (a_5 \cap a_2)$

$a_2 \equiv p_3$ $\quad b_2 \equiv (a_4 \rightarrow a_2)$

$a_3 \equiv (p_2 \rightarrow p_1)$ $\quad c_1 \equiv (b_2 \cap a_3)$

$a_4 \equiv p_0$ $\quad c_2 \equiv (b_1 \rightarrow p_2)$

$a_5 \equiv (p_0 \cap p_1)$ $\quad d_1 \equiv (c_2 \cap a_1)$

$e_1 \equiv (d_1 \cap c_1)$

The computer may now compute truth values for the whole sentence by finding the truth values for the "nested" two-variable sentences.

Another problem involving the partial generation of a truth table is that of finding only those values of the variables which are compatible with the statement of the problem. As an example of how a problem of this sort could be handled on a digital computer, the following illustration is presented. A part is manufactured by three machining operations, one in department I, one in department II, and one in department III. Department I has five kinds of machines which can be used, II has six, and III has three. However, because of scheduling problems and property changes in the part caused by using a particular machine in a previous department, there are restrictions on the sequence of machines which may be used to produce an acceptable part. The departments and machines are represented below:

Stated symbolically, the manufacturing process is:

$$\text{I} \cap \text{II} \cap \text{III} \equiv (A \cup B \cup C \cup D \cup E) \cap (F \cup G \cup H \cup I \cup J \cup K) \cap (L \cup M \cup N)$$

(Which reads, the combination of processes in departments I, II, and III is equivalent to the use of A or B or C or D or E; and the use of F or G or H or I or J or K; and the use of L or M or N.)

A series of conditions in the form $A \rightarrow B' \cap C' \cap D' \cap E'$ are written, their total effect being that the process is limited to the use of one and only one machine per department. The restrictions on the use of particular machines in sequence have been determined from shop studies, and are as follows: [3]

$A \rightarrow I' \cap J' \cap K' \cap L'$

$B \rightarrow J' \cap K' \cap N'$

$C \rightarrow F' \cap J' \cap M'$

$D \rightarrow F' \cap H' \cap M'$

$E \rightarrow F' \cap G' \cap I' \cap J' \cap K' \cap L' \cap M'$

$F \rightarrow C' \cap D' \cap E' \cap N'$

$G \rightarrow E' \cap M' \cap N'$

$H \rightarrow D'$

$I \rightarrow A' \cap E' \cap L' \cap M'$

$J \rightarrow A' \cap B' \cap G' \cap L'$

[3] These restrictions are from a problem in (4).

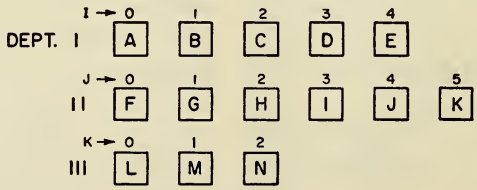


FIGURE 4

$K \rightarrow A' \cap B' \cap E' \cap M'$
$L \rightarrow A' \cap E' \cap H'$
$M \rightarrow C' \cap D' \cap E' \cap G' \cap I' \cap K'$
$N \rightarrow B' \cap F' \cap G'$

The restrictions generate a truth table, which can be stored in the memory of a computer in the following manner: (See Figure 5.) The *i, j, k,* notation is shown in the representation of the departments and machines above. The computer is then instructed by the program of Figure 6. The computer will then print those sequences of machines on which acceptable parts may be produced.

The application of digital and analog computers to the solution of problems in symbolic logic extends the usefulness of the method greatly. E.g., a problem of the type just discussed could have ten machines in each department, and seven departments without using all the memory of a computer of the IBM 650 type. This represents a total of ten million possible sequences. Although the size of symbolic logic problems solvable on a digital computer is ultimately limited by the memory storage capacity of the computer, the problems lying within this limit are not trivial.

TABLE IJ

J \ I	0	1	2	3	4	5	6	7	8	9	MEMORY LOCATION
0			F	F	F						001
1					F						002
2				F							003
3	F				F						004
4	F	F	F		F						005
5	F	F			F						006

TABLE IK

K \ I	0	1	2	3	4	5	6	7	8	9	
0	F				F						007
1			F	F	F						008
2		F									009

TABLE JK

K \ J	0	1	2	3	4	5	6	7	8	9	
0			F								010
1		F		F		F					011
2	F	F									012

FIGURE 5

In a recent article (6), it has been suggested that automation has been applied in manufacturing only in obvious situations where there is a demand for a high-production item. The author of the article suggests that additional applications for automation exist where a single sub-assembly or part is common to a variety of different items being produced. Further, a minor redesign of similar parts could yield a standard part which could be used on a large number of items. However, the identification of similar parts, or rather, parts which perform similar functional requirements, is a difficult task in a complex industrial situation. Symbolic logic can be applied in these circumstances to reduce the functional requirements of each part to concise logical sentences. Once stated in symbolic form, each functional requirement, which represents a part or a designer's need for a part, can be compared to tabulated functional requirements, which represent standard parts being made or which are available. Consider the following hypothetical example: A company making business machines uses a large variety of springs. The total number of springs used is extremely large, but the quantity of each variety is too small to permit an economical use of automation in manufacture. It is suspected that the large variety used is partly due to the fact that in situations where a number of different springs would satisfy the functional requirements equally well, one spring is specified by the designer, and from that point forward in the design and development procedure, the original functional requirements are forgotten, and only that one spring is considered for the job. To find opportunities for automation, the functional needs of each situation where a spring is required are stated symbolically.

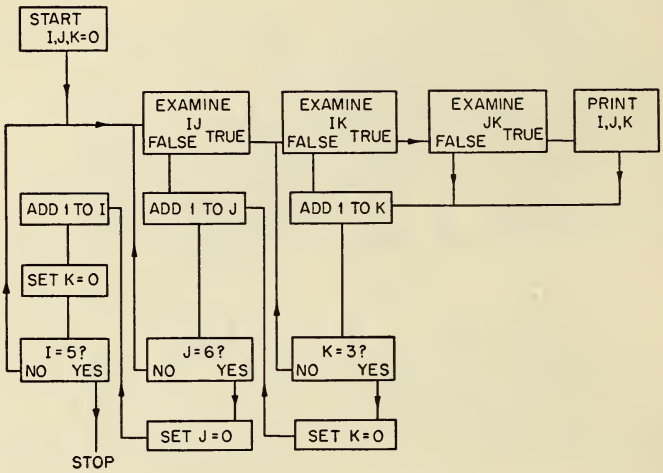


FIGURE 6

$a_1 \equiv$ Spring constant k, 1 ± 1
$a_2 \equiv$ Spring constant k, 3 ± 1

$a_3 \equiv$ Spring constant k, etc., for the entire range of requirements.
$b_1 \equiv$ Length, $0.25'' \pm 0.125''$
$b_2 \equiv$ Length $0.5'' \pm 0.125''$
$b_3 \equiv$ etc., for the entire range of length requirements.
$c_1 \equiv$ etc., for all the classifications of descriptions needed to completely define all the functional requirements of a spring.

The functional requirements of any situation which calls for a spring may now be stated in a logical sentence; e.g., $(a_1 \cup a_2 \cup a_3) \cap (b_3 \cup b_4) \cap (c_5)$, etc. The logical sentences may be coded and punched on IBM cards. The problem then reduces to an ordinary collation procedure. In this manner, all the functional requirements which overlap enough to permit the use of one standard part may be identified; and if the demand for that standard part is sufficient, automated machinery for its production may be economically feasible. In addition, all the standard springs offered by vendors can be described by the same symbols, so that the collator can find what standard commercial springs will meet the functional requirements. Should automation prove impractical, this method will also show where there are opportunities for decreasing the number of different springs in inventory, and possibilities to buy more economical lot sizes.

Descriptions of the operating states of components of complex systems may be written in the same manner that is as ranges of numerical values represented by symbols as in the last example. Using these symbols, the interrelationships of components may be described in logical sentences, and these combined to describe the action of the system when a component's operating state deviates from the performance standard. The logical description of the system also represents the analog of a control mechanism which will alter other components of the system to compensate for that deviation. If the complex system is a business organization, and it is not beyond the realm of reason to describe the operation of an organization of humans symbolically, then the logical description is the analog of a set of rules or policies which govern the behavior of the organization such that it operates in a manner which will insure that the objectives of the organization as a whole are best served.

CONCLUSION

At this point, the reader may wonder if some of the solutions to suggested applications of symbolic logic are merely dressed-up versions of common sense or ordinary engineering procedure, and in reality a solution could be obtained in the same manner by someone who knew nothing about formal logic. The answer to this query lies in the previously stated definition of symbolic logic. It is a means of clearly expressing verbal propositions and their relationships in symbolic form. Someone who knows nothing of formal logic may unknowingly use elementary forms of that method and solve problems. This, however, does not detract from the dignity or usefulness of a formal system of symbolic logic and logical operations. The Egyptian mathemati-

cians mentioned earlier knew little of the methods of a formal algebra, but managed to solve elementary problems. This certainly does not argue against algebra as a formal method, or its dignity as an element of mathematics.

In conclusion, a discussion of the advantages and disadvantages of symbolic logic as applied to practical situations is apropos.

An analysis of a complex situation can be made, and the results are in a concise and unambiguous form. Manipulation of component elements describing the system under consideration are performed in a rigorous manner, and conflicts and redundancies are disclosed and eliminated.

The disadvantages of symbolic logic in a practical situation are those of every model. A model is only an approximate description of a system, and as such, one can neither use it to exactly predict the behavior of the system, nor expect it to be valid when the system changes; i.e., when it is no longer the system for which the model was designed.

To compound this inherent limitation, the translation of complex verbal relationships into symbolic form is not simple. Logicians have grappled with the problem of translation since Boole's original dissertation, and the literature describing techniques is prodigious. The material covered here has been of the most elementary nature.

This logic is two-valued; i.e., each proposition may be either true or false. Probabilistic systems would require an extension of the method to include alternative solutions based on the probability of the stated relationships being true. This could conceivably be accomplished on digital computers by combining the solution of the logical problem with Monte Carlo or other statistical methods (4).

Symbolic logic has been presented as a tool for analysis of a practical situation. Its value lies in its ability to reduce complex problems to clear, concise terms. A knowledge of the subject, including its capabilities and limitations, is a valuable addition to . . . [business] methodology.

REFERENCES

(1) Abbot, Wilton R., "Computing Logical Truth with the California Digital Computer," *Mathematical Tables and Other Aids to Computation*, Vol. 5, No. 35, (July 1951), pp. 115–184.

(2) Birkhoff, Garrett, and MacLane, Saunders, *A Survey of Modern Algebra*, MacMillan and Co., New York, 1953 (with permission of MacMillan and Company).

(3) Boole, George, *An Investigation of the Laws of Thought, on which are Founded the Mathematical Theories of Logic and Probabilities*, MacMillan and Co., London, 1854 (with permission of MacMillan and Company).

(4) Cushen, Walter E., "Symbolic Logic in Operations Research," *Operations Research for Management*, Baltimore, The Johns Hopkins Press, 1954.

(5) Gardner, M., "Logic Machines," *Scientific American*, March 1952, pp. 68–73.

(6) Hurni, M. L., "Increasing the Opportunities for Automaticity," *Modern Materials Handling*, Vol. X, No. 1, January 1955.

(7) Lanning, Walter C., "Boolean Algebra and Its Application in Logical Design," *Sperry Engineering Review*, May–June and July–August 1956.

(8) Pfeiffer, J. E., "Symbolic Logic," *Scientific American,* December, 1950, pp. 22–24.

(9) Strawson, P. F., *Introduction to Logical Theory,* London, Methuen and Co., Ltd., 1952.

IV FACTOR ANALYSIS

✦✦✦✦✦✦✦✦✦✦✦✦✦ FACTOR ANALYSIS: *when to use it*

CHARLES K. RAMOND

Not long ago, sales of a well known food product began to decline swiftly. The agency account group, assembled and asked why, answered by listing factors they agreed were most influential. The list was called "A Factor Analysis of Product X Sales."

Of course this was not factor analysis as the scientist understands it. To him factor analysis is a way of describing the inter-relationships between a large number of variables in terms of a few mathematically derived factors. Most of us need to be reminded, however, that both of these procedures—the listing of possible causes and the mathematical transformation of correlations between them into factors—are essentially similar steps in the scientist's task. Both are preliminary to the testing of hypotheses in experiments.

The scientist's task is to learn what causes what—which stimuli cause which responses? Often he can do this only by experiment: subjects are randomly assigned to groups receiving different treatments or different levels of some independent variable. What follows here is about how the scientist chooses *which* variables to investigate.

The applied scientist has little difficulty in choosing or defining a response measure. It may be "given" by his field of interest or by his employer's. Sales, brand awareness, attitude change and consumer preference are some common response measures in marketing.

His choice of stimulus variables may also be limited by circumstances. If a scientist works for a drug company, he may investigate nothing but different dosages of each new drug developed. If he works for a mail order distributor, his only independent variable could well be the amount of space allocated to each product in the company catalog.

Based on a talk given to the Statistical Methods and Operations Research Discussion Group, The American Marketing Association, March 30, 1961.

More often, however, outside conditions dictate only which response to measure. The result desired is known; the stimulus variables which cause it are either unknown or so numerous as to preclude investigating them all in a single experiment. Indeed, almost every scientist sooner or later faces the awesome fact that there is literally an infinite number of experiments he could do—even if he is interested in predicting only one kind of response.

For most effects we prize have many causes. Often these causes interact in some complex way to produce a response.

So the initial step in many research programs is the identification of possibly relevant stimulus variables. This is not always done most efficiently. The experimental literature of any discipline shows that the scientist's personal whims and temperament usually dictate which independent variables he investigates. His desire to be original leads him to the new and different.

This is especially true of graduate students who seek to make an original contribution. Since much published research is done by such students, the experimental literature, particularly in the social sciences, is fragmented and individualistic. Theoretical systems (or powerful theorists) work to bring order out of this chaos, but where their influence is felt, it is usually controversial.

Perhaps the great scientist differs from merely competent ones mainly in his ability to choose the right experiment to do next. Intuitively he perceives complex inter-relationships, neglected factors, realities veiled by conventional wisdom. In choosing experimental variables to investigate he draws upon these perceptions, takes greater risks, and often finds a relationship no one else suspected could exist. His experiments, therefore, become demonstrations or confirmations of his own genius in observing how the world works.

Most of us must be content with more modest manifestations of our scientific skill. Many scientists, including most graduate students, need a mechanism by which they can simplify and reduce a fantastic number of independent variables to a few which can be investigated. Even the great scientist may profitably rely on such summary procedures in dealing with a complex field.

There are many such procedures. One is factor analysis.

Let us describe this procedure nonmathematically, not telling "how to do it," but rather to show the doer, user or connoisseur of research how it can work to his advantage.

A factor analysis has only one practical purpose: to reduce the number of experiments the scientist must do to impose order on complex subject matter. It helps him (or his manager) decide which experiments should be done next. So many responses are caused by large numbers of stimuli that his factor analysis may be the only way the researcher can reach a conclusion—within his lifetime—about which stimulus variables are most important in producing the response he wishes to predict or control.

We know that correlation coefficients describe the extent to which two meas-

ures of the same thing vary together. When a large number of such variables are considered at once, a convenient way to list all of their inter-relationships simultaneously is in a table of their correlations with each other. Such a table is called a correlation matrix.

TABLE 1

A CORRELATION MATRIX
(From Stoetzel, 1961)

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
B	.21							
C	.37	.09						
D	−.32	−.29	−.31					
E	.00	.12	−.04	−.16				
F	−.31	−.30	−.30	.25	−.20			
G	−.26	−.14	−.11	−.13	−.03	−.24		
H	.09	.01	.12	−.14	−.08	−.16	−.20	
J	−.38	−.39	−.39	.90	−.38	.18	.04	−.24

(There is no need to show the correlations in the other half of the table: they would be the same as their counterparts above. As in a roadmap distance table, where the distance from *A* to *B* equals that from *B* to *A*, correlation *AB* equals correlation *BA*.)

Such a table is the grist for a factor analysis. The latter seeks to reproduce the correlations between many variables by assuming the existence of a few factors which account for them all.

For example, if in the above table *A* stood for height and *B* for weight, and if these were correlated .91 instead of .21, we would lose little information about the individuals described if we dropped one of these two variables and simply let it be represented by the other. This process can be extended.

What is meant by one variable accounting for another? Suppose that the variables in the table refer to performance in a series of track events, and that *A* through *G* stand for running speed in dashes of 60, 100, 220, etc. yards. These scores would be highly correlated and one factor of running speed could predict all of them fairly well.

This factor is an abstraction that can be calculated from the intercorrelation shown above. It is not some additional measure which we introduce in the hope that it will be observed to correlate with all the running measures. It is rather a name for the regularity we can observe in the data we already have.

As a matter of fact, the only way we can see the existence of this abstraction is to tabulate the correlations between it and the original experimental variables. Such correlations are called factor loadings and a typical table of them is shown below.

TABLE 2

FACTOR LOADINGS
(From Stoetzel, 1961)

Item	*Factor I*	*Factor II*	*Factor III*
J	0.64	0.02	0.16
D	0.50	−0.06	−0.10
F	0.46	−0.24	−0.19
G	0.17	0.74	0.97
E	−0.29	0.66	−0.39
H	−0.29	−0.08	0.09
B	−0.49	0.20	−0.04
C	−0.52	−0.03	0.42
A	−0.60	−0.17	0.14

Read as follows: factor I is correlated .64 with performance in the 60 yard dash, .50 with performance in the 100 yard dash, etc. The computation procedures of factor analysis insure that the first factor to emerge will be that which accounts for the most variance in the original measures correlated. The second factor accounts for the next most and so on. The analysis is stopped when the last factor generated accounts for less variance than any independent variable itself can predict. Factor analyses of complex data easily produce more than six factors before this point is reached.

We need not describe the computations involved in going from the correlation matrix to the table of factor loadings. Suffice it to say that these involve the solving of many equations simultaneously, one for each correlation in the table. One such equation would look like this:

r_{ab} = (*A*'s loading on factor I) × (*B*'s loading on factor I) + (*A*'s loading on factor II) × (*B*'s loading on factor II) + · · · · · + (*A*'s loading on factor *N*) × (*B*'s loading on factor *N*)

The idea is to find the loadings on the above equation which best reproduce the correlations between *A* and *B*, as well as all other correlations in the table. This would be an extremely time-consuming task for an individual using the calculator. Happily standard computer programs exist to perform this busy work.

Less happily, the mathematics of factor analysis is so intriguing to certain scientists that they sometimes invest the factors with explanatory power. Let us make a distinction: descriptions merely reduce the number of observations without losing much of the information they contain, while explanations are statements of repeatable cause-effect links and thus require the ability to predict successfully. The factors which result from the factor analysis have no such predictive power—

they are merely a few variables which can do the same descriptive job as the larger number of variables from which they were extracted.

For efficiency of description they are peerless. For showing underlying causes of any effects they are only a first step.

The perilous point for a factor analyst comes at the moment when he first sees the loadings of the original independent variables on each factor. Quickly he arranges the independent variables in order to correspond to their loadings on factor I. A pattern seems to emerge. He is unable to resist the temptation to name the factor.

For example, he sees that the highest loadings on factor I are those of all the speed events, while the lowest loadings are those of the events requiring no rapid movement. Why not call this the speed factor? The danger lies in confusing this *naming* with the demonstration of true cause and effect. No factor from a factor analysis ever caused anything—except possibly confusion between description and explanation.

We explain by showing cause and effect. We show causality by experimentation. The best use that can be made of the factors from a factor analysis is to define them operationally, thereby to use them as treatments in an experiment. Instead of laboriously investigating a large number of independent variables in several experiments, we may simply examine various operational definitions of the factor which seems to underlie them all.

Recall for a moment the only evidence we have for the "existence" of a factor: the column of loadings showing the degree to which it correlates with each of the independent variables. The pattern we see in these newly ordered variables leads us to postulate a factor and to name it. This naming is a subjective process. Different individuals may see different patterns.

One analysis of French consumer preferences (Stoetzel, 1960) ranked nine liquors by their loadings on the first factor as follows: Liqueurs, Kirsch, Mirabelle, Rum, Marc, Whiskey, Calvados, Cognac and Armagnac. Stoetzel chose to label this factor "sweet-to-strong." Perhaps you agree that Liqueurs and Kirsch taste sweet, and Cognac and Armagnac taste strong. But as Stoetzel aptly noted, passing from the mathematical to the psychological is indeed delicate.

Passing from the abstract to the concrete is equally delicate. How may we operationally define the factor sweet-to-strong in order to test its influence on consumer preference experimentally? One way might be in terms of the proportion of sugar in an alcoholic beverage. But this excludes the notion of strength. Instead we might define the sweet-to-strong factor as the ratio of sugar to alcohol in the beverage.

We could use a variety of such operational definitions. How has this reduced the number of experiments we might do? Only by focusing our attention on a few variables, all of which can be defined so as to reflect the most important factor indicated by the column of loadings. Without the factor analysis, we might not have plumbed this one small area so deeply or so soon.

Then the proof of the pudding is in

the experiment. Our hypotheses are clear. Some operationally definable factor which ranks liquors as above seems to cause differences in consumer preferences for these liquors. If and when the experiment shows or fails to show the effect of this factor may we accept or reject this hypothesis.

So factor analysis is a way to generate more fruitful hypotheses. It does not demonstrate causality, though some investigators behave as if it did. Too many studies stop after a factor analysis without going on to do the experiments it suggests.

So much for the rationale of factor analysis. Let us now consider the situations in which this technique is particularly useful.

Clearly factor analysis is no tool for a researcher who has already decided which hypotheses he wishes to test in his next experiment. Rather it is more useful in situations where good hypotheses are non-existent or too plentiful. We should either have no idea whatsoever of which independent variables cause the response, or be able to list dozens or hundreds of such variables which may be influential.

In both of these situations, experiments are temporarily delayed. In the first, we have no idea where to begin; in the second we have too many.

An example of the first case was previously mentioned, Stoetzel's factor analysis of liquor preferences. Apparently he was unable or unwilling to guess which characteristics of liquors caused them to be preferred over one another. He therefore asked some 2,000 Frenchmen to rank these liquors and analyzed rank correlations between liquors to observe which factors, if any, seemed to emerge.

This situation often prevails in the initial stages of advertising research when one is trying to decide what to say. Judging from the name he gave to the first factor which emerged, Stoetzel might well have decided to experiment with themes emphasizing the liquor's sweetness or strength.

Once we have decided what to say, the next decision is usually how to say it. Here the situation is often just the opposite. Instead of the absence of possible causes, we are often confronted with more than we could possibly investigate.

An excellent example is the factor analysis of recognition scores of 137 advertisements in the February 1950 issue of the *American Builder* (Twedt, 1958). Most of us here could list without hesitation a few dozen characteristics of industrial advertisements which we suspect might contribute to its recognition by a reader. Twedt measured 34 such characteristics of each ad. These included such mechanical variables as size, width-height ratio, number of colors, square inches of illustration, point size of main body copy, and so on. He also measured a number of content variables including Flesch readability scores, number of words in headlines, number of product facts listed, number of pictures of product in use, and the like.

To begin his analysis he selected from these 34 variables those 19 which correlated significantly with the readership criterion. Intercorrelations among these 19 plus the criterion were incorporated in a 20 x 20 correlation matrix,

which was factor analyzed by Thurstone's complete centroid method. The analysis was stopped with the sixth factor on which the product of the two highest loadings was just equal to the standard error of the original correlation between these two variables.

Factor one had the highest loading on square inches of illustration, number of pictures showing the product in use and number of colors, and was called the pictorial-color factor. The second factor had the highest loadings on ad size, number of product benefits, square inches, and number of words in the advertisement. Somewhat more cautiously the author called this the size factor. The third factor had high loadings on largest type size, largest type used for product identification, number of type faces and point size of main body copy. This factor was called typographic size and variety.

In similar fashion, the remaining three factors were called information, field and previous advertising schedule. Collectively the six factors accounted for about two-thirds of the observed variance in readership scores. The pictorial color factor alone accounted for 41 per cent and with the size factor 53 per cent of this variance.

In contrast to the Stoetzel report which ended at the point of reporting the factor loadings, Twedt went on to test those independent variables which the factor analysis suggested were purest and most influential. No true experimental design was followed; rather, Twedt calculated the multiple correlation coefficient between these suggested variables and the readership criterion.

The three variables were size of advertisement, number of colors and square inches of illustration. They correlated .77 with readership of the same ads in the *American Builder*. Obviously the value of this finding is in its applicability to other data. Twedt therefore used this same multiple regression formula to predict the readership scores of ads in five other magazines for which actual readership scores were available. The correlations between these actual scores and those predicted by the regression formula ranged from .58 to .80 depending on the magazine and averaged .71. Considered as validity coefficients, these correlations are rather high. The point to bear in mind, however, is that the regression formula by which the readership scores were predicted was only one of thousands that could have been chosen. The investigator was led to one of the best in rapid fashion by means of the factor analysis.

This paper tries to show—without mathematics—what factor analysis is and when it should be used. Factor analysis is essentially a tool for helping decide which experiment to do next and it finds its most profitable application in the cases where the experimental variables are unknown or too profuse. Remember: in determining the causes of any effect, the factor analysis can never supplant the experiment. It can only lead us to the right experiment sooner.

REFERENCES

Stoetzel, Jean. A Factor Analysis of the Liquor Preferences of France Consumers. *Journal of Advertising Research*, Vol. 1, No. 2, December 1950.

Twedt, Dik Warren. Why Advertising Readership Varies. In *Motivation and Market Behavior* (Ferber and Wales, Editors). Homewood, Illinois: Richard D. Irwin, Inc., 1958.

✦✦✦✦✦✦✦✦✦✦✦✦ FACTOR ANALYSIS: *how it's done*

GWYN COLLINS

Briefly factor analysis is a technique for representing a large number of measurements, each made on many objects or persons in terms of some smaller number of artificial measurements. For instance, the results of twenty different tests of mechanical ability may be factor analyzed and from the analysis we might find that from just four indexes, or factors, for each person tested, we could reproduce all twenty scores with a high degree of accuracy. This kind of thing is the basis of factor analysis. Unfortunately for simplicity, there are many ways of doing it and something needs to be said about them and their points of difference.

Factor analysis is the collective name given to all these procedures for deriving factors from original measurements. Its central idea was suggested 75 years ago as a means of isolating the fundamental characteristics of convicted criminals; its techniques were first introduced almost 60 years ago by a British engineer, Charles Spearman, in an article on intelligence in a psychological journal. Since then, its practice has been almost entirely confined to psychology, especially to the field of personality measurement. Spearman's paper was itself entitled "General Intelligence, Objectively Determined and Measured." It presented the technique as a means of resolving disputes about the meaning of intelligence, but the method he presented, and later elaborations of it, have served to feed controversy in this very area right up to the present time.

Professional mathematicians and statisticians have had little to say about factor analysis; mathematicians because the basic procedures have long been known and are relatively trivial, statisticians because the statistical problems of factor analysis such as that of estimating sampling errors, have appeared more or less beyond solution. To statisticians generally, factor analysis has exerted very little appeal; to some it is a dirty word.

The raw material of factor analysis is a table of original measurements. Suppose that we have many individuals a, b, c—to n and that we have measured each of them in various ways A, B, C—

Journal of Advertising Research, *September 1961, 1:5, 28–32.*

to M, all different. We can record the results in a matrix or rectangular table of numbers, like this:

	INDIVIDUALS					
Tests	a	b	c	d	...	n
A	X_{Aa}	X_{Ab}	X_{Ac}	X_{Ad}		
B	X_{Ba}	X_{Bb}	X_{Bc}			
C						
D						
.						
.						
.						
M	X_{Ma}	X_{Mb}	X_{Mc}	...	...	X_{Mn}

where X_{Bc}, for instance, denotes the measurement B taken on individual c. This, of course, is just a straightforward tabulation of the result of each of the measurements. Since they may be taken on a very large number of people, this is normally a huge matrix.

FACTORS ARE FEWER

Now the object of factor analysis is to find certain new composite dimensions, or factors, say α, β, γ, δ etc., [alpha, beta, gamma, delta etc.,] fewer in number than the original measurements, uncorrelated with one another and which contain all the information provided by them. Each individual would have a score on each of these factors though it would not be open to direct observation; instead it would have to be inferred from the original measurements. The result of measurement A taken on individual a has been written in the matrix of original measurements as X_{Aa}. Analogously we can write his hypothetical score on the artificial dimension α as $X_{\alpha a}$, on the dimension β as $X_{\beta a}$ and so on. The problem is to find new dimensions such that any individual's score can be reconstituted by adding together some proportion of each of his hypothetical scores. Moreover, these proportions must be the same for every individual.

These proportions obviously have great importance and are appropriately called *factor loadings*. Thus if X_{Aa} is written in terms of factor scores as:

$$X_{Aa} = pX_{\alpha a} + qX_{\beta a} + rX_{\gamma a} + \text{etc.},$$

then we would write X_{Ab} as:

$$X_{Ab} = pX_{\alpha b} + qX_{\beta b} + rX_{\gamma b} + \text{etc.},$$

where p, q and r, the proportions of the factors α, β and γ that are used to reconstitute the measurement A, are known as the factor loadings of measurement A on the factors α, β and γ.

An important point is that factor analysis only attempts to find factors which can be added in this simple way to get back to the original measurements. It is conceivable that the original measurements could be obtained by some more complicated way of putting factor scores together, but this possibility is not encountered in present-day factor analysis. What we have described is an ideal which no procedure achieves.

TWO APPROACHES

The attempt to reconstruct original measurements from a smaller number of factors is in practice rarely, if ever, completely successful. Usually it can only be done approximately and this offers the factor analyst a choice. He

can either go on extracting factors until he has enough to reconstitute the original measurements almost perfectly, or he can decide at the outset that he will look for those two, three, or however many factors he prefers, which do the best job of re-creating the original measurements. These two approaches are really fundamentally different, though the difference is usually ignored. The first approach, that of taking out as many factors as necessary, starts with the observed data and seeks to represent it in another way in terms of factors uncorrelated with one another. The second approach, that of finding the "best" p factors (where p is some number decided upon beforehand) is an attempt to fit a mathematical model to the observed data. The first approach is more accurately described as *component analysis*, and the term *factor analysis* should perhaps be reserved for the second approach. The confusion between these two basic methods has arisen because the arithmetical routines of both are closely similar. No attempt is made in the remainder of this article to unscramble this well established confusion; both will usually be referred to indiscriminately as factor analysis.

Factor analysts start their work by summarizing the information from the matrix of original measurements by calculating the correlation coefficients between each pair of measurements. This is a starting point in picking out that which is common to the various measurements. The correlation coefficients are then written in the form of a matrix like this:

	A	B	C	D	—	—	M
A	r_{AA}	r_{AB}	r_{AC}	r_{AD}	—	—	—
B	r_{BA}	r_{BB}	r_{BC}	r_{BD}	—	—	—
C	r_{CA}	r_{CB}	r_{CC}	r_{CD}	—	—	—
D	r_{DA}	r_{DB}	r_{DC}	r_{DD}	—	—	—
—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—
M	—	—	—	—	—	—	—

This matrix is obviously square for there are as many rows as there are columns and usually, since there are many less measurements than people, it is a much smaller matrix than that of the original measurements. Another point is that the part of the matrix above the diagonal running from top left to bottom right (called the *principal diagonal*) is reflected on the other side of it because, for instance, the correlation between C and A, r_{CA} is equal to the correlation between A and C, r_{AC}. Another point is that the principal diagonal consists of terms like r_{AA}, r_{BB}, r_{CC} and so on. Now it might be thought that these entries must necessarily all equal 1, for the correlation of any measurement with itself should be perfect. In this case, however, we are dealing with the *hypothetical* correlation which would be found between two actual measurements. Since there are always at least errors in measurement, we might expect such correlations to be high but not necessarily perfect. Just how the spaces in the main diagonal should be filled is an important issue in factor analysis.

Earlier, in discussing the matrix of original measurements, factor analysis was described as an attempt to express each of the measurements in terms of

some linear combination of a lesser number of them where the new artificial measurements, or factors, are uncorrelated. If this were possible then we should describe the measurements as *linearly dependent:* if it is impossible we describe the measurements as *linearly independent.* Now if all the M measurements of the matrix of original measurements are linearly independent, we describe that matrix as having *rank* M. If, however, any one of the measurements could always be obtained by a combination of the others, the matrix would be said to have rank (M — 1) at most. In short, the rank of a matrix is the greatest number of linearly independent rows (or columns) to be found in it. Using these terms, we can say that factor analysis aims to express the matrix of original measurements as a matrix of some reduced rank. Since it is easy to show mathematically that the rank of the correlation matrix is always the same as that of the original matrix (if all elements of the principal diagonal are made equal to one) from which it is derived, the same thing can be said of the correlation matrix.

FACTOR VS. COMPONENT ANALYSIS

At this point, however, it is worthwhile to make some distinction between factor analysis and component analysis. The factor analyst is interested in actually reducing the rank of the correlation matrix to the number of factors he thinks is appropriate in his mathematical model. The component analyst, on the other hand, is not interested in reducing the rank of the correlation matrix. He is interested instead in obtaining factors which are not correlated with one another such that the first will explain as much as possible of the original measurements, the second will explain as much as possible of that left unexplained and so on.

Now the factor analyst wishing to reduce the rank of the correlation matrix has one useful tool. It has been pointed out that the elements of the principal diagonal of that matrix are more or less undefined. If appropriate numbers are put in, the rank of the matrix can, under very general conditions, be considerably reduced. A correlation matrix of rank 10, for example, can be reduced to rank 6. That is, we can reduce the results of 10 measurements to scores on 6 factors, whatever the original scores are. In fact the correlations actually found between the different measurements may help to reduce the rank of the matrix still further.

There are, however, two points of view about these diagonal entries, which factor analysts call *communalities.* One point of view expressed by a statistician is:

> Where the [communalities] are completely unknown one method of approach has been to regard them as being at choice; and in particular to assume that they are such as to minimize the number of factors. In general, this seems to assume, on Nature's part, a much more indulgent behavior than we have any right to expect, but it is interesting to see what happens in such cases. (Kendall, 1957, p. 43.)

The other point of view, often expressed by psychologists, is that the communalities are unique numbers which represent that portion of each measurement which correlates with other measurements under consideration; they are not observed, but can be computed. This, however, seems to be a way of describing an aspect of the mathematical model of factor analysis rather than a statement of observable fact.

There is another useful way of looking at the meaning of the rank of a matrix. Suppose, for instance, that we have only two measurements for each of a number of individuals. We could, if we chose, represent the pair of measurements for each person as points on a graph. The distance of the point from one axis could be used to represent one measurement; the distance from the other axis could stand for the other measurement. Points, then, represent persons and because their positions depend upon two measurements they are spread over a two dimensional space —a plane. If we had three measurements for each person we would, of course, need three different dimensions in which to represent them, and so on.

To this rule of n dimensions to represent n different measurements there is an interesting exception. If, for example, we have two measurements for each person and the second is, say, always three times the first, then the points representing persons will always fall on one straight line. Because both measurements can be expressed in terms of only one of them, only one dimension is necessary to represent both. Similarly, if three measurements are such that they can be expressed in terms of any two of them, all the points will lie in two dimensions and, in general, n different measurements not linearly independent can be represented in less than n dimensions.

If the original measurements were expressed in terms of factors uncorrelated with one another, we should be able to represent them diagrammatically in a space with as many dimensions as there are factors. In the graph each measurement would be denoted by a point whose distances from the axes (each of which represents a factor) would represent factor loadings. If lines are drawn to join each of these points to the origin the angle between any pair of them can be shown to be a simple function of the correlation coefficient between the measurements they represent (the correlation coefficient is the cosine of the angle). Observed correlations can be represented completely by angles between pairs of lines. In the case of three dimensions we may imagine a series of spokes sticking out in all directions from some fixed point as representing the original measurements—this, of course, is an illustration of the original measurements represented in a three dimensional space, that is by three factors. Now the factors are inferred from the correlation coefficients while the correlation coefficients are derived from observed data. We cannot choose our correlation coefficients; we must put up with those we find. The factors, however, can be chosen in any way that we please, as long as they do the job of representing the original measurements.

Returning to our three dimensional example, the three factors can be represented as three axes perpendicular to one another with their origin at the common point of the lines standing for the original measurements. Now the only observable data expressed in the diagram are the correlation coefficients and these are denoted by the angles between the lines. These angles are, then, the only invariant part of the diagram. But it is clear that if the whole configuration of lines representing the original measurements were rotated with its origin held fixed, the lines would then occupy different positions with respect to the axes and would have to be described differently. In other words, when we represent a correlation matrix in terms of a number of factors, there are an infinite number of ways of doing it. Each of these corresponds to a different rotation of the factor axes relative to the lines representing the original measurements.

SIMPLE STRUCTURE

This is a problem which has exercised psychologists greatly. A number of criteria have been suggested for the "best" selection of axes. The most venerated of these is one proposed by Thurstone and called "simple structure." The use of "simple structure" is no more than the factor analytic expression of the principle of parsimony. Essentially, it is that each variable should be represented by as few factors as possible. The rules given by Thurstone for obtaining "simple structure" demand a certain amount of judgment from the analyst. In general "simple structure" means that the axes will be so chosen that a large number of the points representing the original measurements will be near the origin and many of them will be close to one axis or another.

More recently there have been a number of attempts to specify what is meant by "simple structure" in analytical terms. Two well known ones are referred to as the Varimax and Quartimax methods. Roughly speaking the Quartimax method is a method of rotating the axes so that each measurement is described in terms of as few factors as possible. The Varimax method, on the other hand, obtains a rotation of the factor axes so as to minimize the number of measurements in which any one factor occurs.

The various procedures that have been used in factor analysis very often provide different approximations to the same factor solution. Usually they can be translated into one another mathematically or, where their mathematical models differ, their differences are, of course, the result of different initial assumptions made by the analyst. In the early days of factor analysis, this was not so clear and disputes between rival analysts were heated and sometimes abusive. One review of rival methods put it this way:

> Factor theory may be defined as a mathematical rationalization. A factor-analyst is an individual with a peculiar obsession regarding the nature of mental ability or personality. By the application of higher mathematics to wishful thinking, he always proves his original fixed idea or compulsion was right or necessary. In the proc-

ess he usually proves that all other factor-analysts are dangerously insane, and that the only salvation for them is to undergo his own brand of analysis in order that the true essence of their several maladies may be discovered. Since they never submit to this indignity, he classes them all as hopeless cases, and searches about for some branch of mathematics which none of them is likely to have studied in order to prove that their incurability is not only necessary but also sufficient. (Cureton, 1939, p. 287).

Nowadays three procedures seem to share public favor. One of them, cluster analysis, is not a factor analytic method at all. It offers an insight into the way measurements group together in terms of their correlation coefficients. A far simpler procedure than factor analysis it is regarded by many as providing an approximation to it.

PRINCIPAL AXES METHOD

The method of factor analysis which has always seemed mathematically preferable is that known as the method of principal axes—this is in fact a component analysis. It achieves a unique resolution of the original measurements into factors. No subjective judgment is called for at any stage of the method. In this method the first axis is selected so that the sum of the squares of the distances of points from the axis is minimized. Successive axes, each perpendicular to the preceding axes, are chosen so as to minimize the squares of the distance of the points from the new axis. Factor loadings on each successive axis become smaller and smaller until the number of factors reaches the rank of the correlation matrix. Usually, the factor loadings on the first few factors can reproduce the correlation matrix very well and the analysis is concluded when such a stage has been reached. Thus the method, a true component analysis, is sometimes used to select a limited number of factors in the same way as a true factor analytic method might.

CENTROID METHOD

The computations required by the method of principal axes are extremely laborious. The most popular method of factor analysis, and one which demands much less labor is the *centroid method,* proposed by Burt (1917) and developed by Thurstone (1931), which was originally intended as an approximation to the principal axes factor solution. Its solution, however, is not unique and some subjective judgment is called for when each set of factor loadings is computed. The *centroid method* selects the first factor axis to pass through the center of gravity of points representing the original measurements. The factor loadings of each measurement on this factor will go a long way toward reproducing the correlation matrix. The difference between the elements of the correlation matrix and those of the attempted reproduction of it, using the first centroid factor, is now regarded as a new correlation matrix and is attacked in the same way. To carry out the next stage, however, it is necessary to alter from positive to negative, or vice versa, all the correlation coefficients involving certain measurements. This selection of measurements for "reflection" of sign, however, is largely subjective.

Currently the principal axes method is gaining favor. Its relative mathematical purity makes it generally desirable and high speed computers have at last made it quite practicable. With the ascendancy of this method the most important questions nowadays in factor analysis are not how it should be done but what it means—and whether it should be done at all.

REFERENCES

Burt, Cyril. *The Distribution and Relations of Educational Abilities.* London: P. S. King & Son, 1917.

Cureton, E. E. The Principal Compulsions of Factor Analysis. *Harvard Educational Review,* Vol. 9, 1939, pp. 287–295.

Harmon, Harry F. *Modern Factor Analysis.* Chicago: University of Chicago Press, 1960.

Kendall, M. G. *A Course in Multivariate Analysis.* London: Charles Griffin, 1957.

Ramond, Charles K. *Factor Analysis in Advertising Research.* Unpublished speech to the Statistical Methods and Operations Research Discussion Group, American Marketing Association, March 30, 1961.

Thurstone, L. L. Multiple Factor Analysis. *Psychological Review,* Vol. 38, 1931, pp. 406–427.

B

Tools for Coping with Variability

I PROBABILITY

PROBABILITY

WARREN WEAVER

Probability is the very guide of life.
—Cicero, *De Natura*

Over three centuries ago some gamblers asked the great Italian scientist Galileo why a throw of three dice turns up a sum of 10 more often than a sum of nine. In 1654 the Chevalier de Mere—another gambler—asked the French mathematician and philosopher Pascal why it was unprofitable to bet even money that at least one double six would come up in 24 throws of two dice. This problem of de Mere really started off the mathematical theory of probability, and the end is not yet in sight.

Probability theory has now outgrown its disreputable origin in the gaming rooms, but its basic notions can still be most easily stated in terms of some familiar game.

Scientific American, *October 1950, 44–46.*

When you toss a die—one carefully made, so that it is reasonable to believe that it is as likely to land on one of its six faces as on any other—a gambler would say that the odds against any specified number are five to one. A mathematician defines the probability to be one-sixth. Suppose we ask now: What is the probability of getting a three and a four in one roll of two dice? For convenience we make one die white and one red. Since any one of six results on the white die can be paired with any one of six results on the red die, there is now a total of 36 ways in which the two can land—all equally likely. The difference in color makes it clear that a red three and a white four is a different throw from a white three and a red four. The probability of throwing a three and a four is the ratio of 2—the number of favorable cases—to 36, the total number of equally likely cases; that is, the probability is 2/36, or 1/18.

What is the probability of throwing a sum of seven with two dice? An experienced crapshooter knows that seven is a "six-way point," which is his way of saying that there are six favorable cases (six and one, one and six, three and four, four and three, five and two, two and five). So the probability of throwing a sum of seven with two dice is 6/36, or 1/6.

In general, the probability of any event is defined to be the fraction obtained by dividing the number of cases favorable to the event by the total number of equally likely cases. The probability of an impossible event (no favorable cases) obviously is 0, and the probability of an inevitable or certain event (all cases favorable) is 1. In all other cases the probability will be a number somewhere between 0 and 1.

Logically cautious readers may have noticed a disturbing aspect of this definition of probability. Since it speaks of "equally likely," i.e., equally probable, events, the definition sits on its own tail, so to speak, defining probability in terms of probability. This difficulty, which has caused a vast amount of technical discussion, is handled in one of two ways.

When one deals with purely *mathematical probability,* "equally likely cases" is an admittedly undefined concept, similar to the theoretical "points" and "lines" of geometry. And there are cases, such as birth statistics for males and females, where the ordinary concept of "equally likely cases" is artificial, so that the notion must be generalized. But a logically consistent theory can be erected on the undefined concept of equally likely cases, just as Euclidean geometry is developed from theoretical points and lines. Only through experience can one decide whether any actual events conform to the theory. The answer of experience is, of course, that the theory does in fact have useful application.

The other way of avoiding the dilemma is illustrated by defining the probability of throwing a four with a particular die as the actual fraction of fours obtained in a long series of throws under essentially uniform conditions. This, the "frequency definition," leads to what is called a *statistical probability.*

On the basis of the mathematical definition of probability, a large and

fascinating body of theory has been developed. We can only hint here at the range and interest of the problems that can be solved. Two rival candidates in an election are eventually going to receive m and n votes respectively, with m greater than n. They are sitting by their radios listening to the count of the returns. What is the probability that as the votes come in the eventual winner is always ahead? The answer is $m - n/m + n$. A storekeeper sells, on the average, 10 of a certain item per week. How many should he stock each Monday to reduce to one in 20 the chance that he will disappoint a customer by being sold out? The answer is 15. Throw a toothpick onto a floor whose narrow boards are just as wide as the toothpick is long. In what fraction of the cases will the toothpick land so as to cross a crack? The answer is $2/\pi$, where π is the familiar constant we all met when we studied high-school geometry. A tavern is 10 blocks east and seven blocks north of a customer's home. If he is so drunk that at each corner it is a matter of pure chance whether he continues straight or turns right or left, what is the probability that he will eventually arrive home? This is a trivial case of a very general "random walk" problem which has serious applications in physics; it applies, for example, to the so-called Brownian movement of very small particles suspended in a liquid, caused by accidental bumps from the liquid's moving molecules. This latter problem, incidentally, was first solved by Einstein when he was 26 years old.

There are laws of chance. We must avoid the philosophically intriguing question as to why chance, which seems to be the antithesis of all order and regularity, can be described at all in terms of laws. Let us consider the Law of Large Numbers, which plays a central role in the whole theory of probability.

The Law of Large Numbers has been established with great rigor and for very general circumstances. The essence of the matter can be illustrated with a simple case. Suppose someone makes a great many tosses of a symmetrical coin, and records the number of times heads and tails appear. One aspect—the more familiar aspect—of the Law of Large Numbers states that by throwing enough times we can make it as probable as desired that the ratio of heads to total throws differ by as little as one pleases from the predicted value 1/2. If you want the ratio to differ from 1/2 by as little as 1/100,000, for example, and if you want to be 99 per cent sure (i.e., the probability = .99) of accomplishing this purpose, then there is a perfectly definite but admittedly large number of throws which will meet your demand. Note that there is no number of throws, however large, that will really *guarantee* that the fraction of heads be within 1/100,000 of 1/2. The law simply states, in a very precise way, that as the number of experiments gets larger and larger, there is a stronger and stronger tendency for the results to conform, *in a ratio sense,* to the probability prediction.

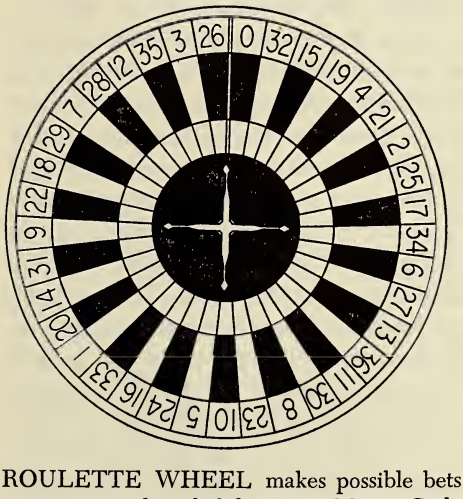
This is the part of probability theory that is vaguely but not always properly understood by those who talk of the "law of averages," and who say

that the probabilities "work out" in the long run. There are two points which such persons sometimes misunderstand.

The first of these relates to the less familiar aspect of the Law of Large Numbers. For the same law that tells us that the *ratio* of successes tends to match the probability of success better and better as the trials increase also tells us that as we increase the number of trials the *absolute number* of successes tends to deviate more and more from the expected number. Suppose, for example, that in 100 throws of a coin 40 heads are obtained, and that as one goes on further and throws 1,000 times, 450 heads are obtained. The *ratio* of heads to total throws has changed from 40 per cent to 45 per cent, and has therefore come closer to the probability expectation of 50 per cent, or 1/2. But in 100 throws the absolute number of heads (40) differs by only 10 from 50, the theoretically expected number, whereas in 1,000 throws, the absolute number of heads (450) differs by 50, or five times as much as before, from the expected number (500). Thus the ratio has improved, but the absolute number has deteriorated.

The second point which is often misunderstood has to do with the independence of any throw relative to the results obtained on previous throws. If heads have come up several times in a row, many persons are inclined to think that the "law of averages" makes a toss of tails now rather more likely than heads. Granting a fair, symmetrical coin, this is simply and positively not so. Even after a very long uninterrupted run of heads, a fair coin is, on the next throw, precisely as likely to come up heads as tails. Actually the less familiar aspect of the Law of Large Numbers already mentioned makes it likely that longer and longer uninterrupted sequences of either heads or tails will occur as we go on throwing, although the familiar aspect of the same law assures us that, in spite of these large absolute deviations, the ratio of heads to tails is likely to come closer and closer to one half.

All of these remarks, of course, apply to a series of *independent* trials. Probability theory has also been most fruitfully applied to series of dependent trials—that is, to cases, such as arise in medicine, genetics, and so on, where past events do influence present probabilities. This study is called the probability of causes.



ROULETTE WHEEL makes possible bets against several probabilities. At Monte Carlo red once came up 32 times in a row. This probability is: $1/(2)^{32}$, or about one in 4 billion.

Suppose we have a covered box about which we know only that it con-

tains a large number of small colored balls. Suppose that without looking into the box we scoop out a handful and find that one third of the balls we have taken are white and two thirds red. What probability statements can we make about the mixture in the box?

This schematic problem, which sounds so formal and trivial, is closely related to the very essence of the procedure of obtaining knowledge about nature through experimentation. Nature is, so to speak, a large closed box whose contents are initially unknown. We take samples out of the box—*i.e.*, we do experiments. What conclusions can be drawn, how are they to be drawn and how secure are they?

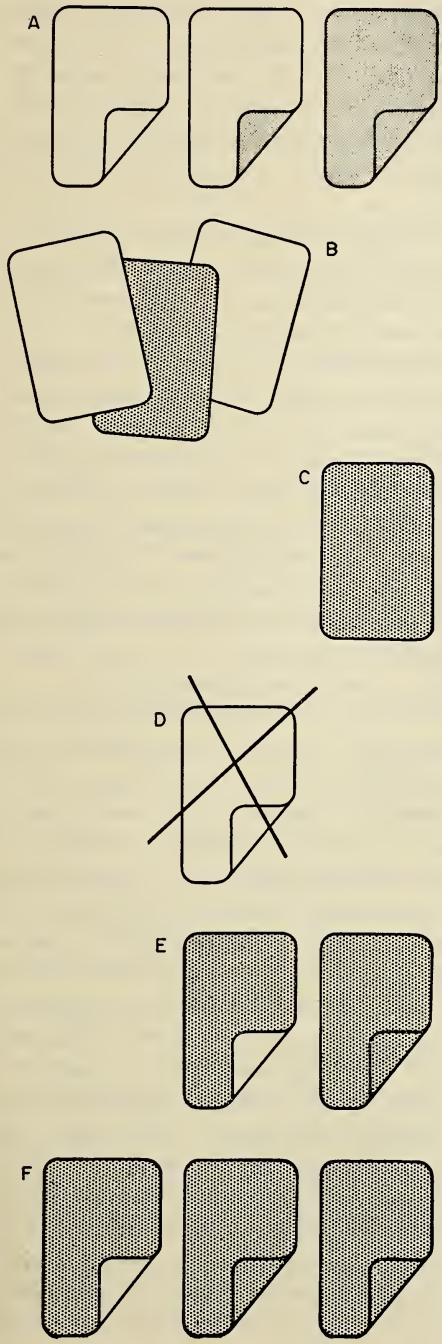
This is a subject which has caused considerable controversy in probability, and in the related field of statistics as well. The problem of the balls as stated above is, as a matter of fact, not a proper problem. The theorem of probability theory which applies here (it is known as Bayes' theorem, and it was first developed by a clergyman) makes clear just how the experimental evidence of the sample justifies one in changing a previously held opinion about the contents of the box; but the application of the theorem requires you to have an opinion prior to the experiment. You cannot construct a conclusion concerning the probability of various mixtures out of the experiment alone. If many repeated experiments continue to give the same indication of one third white and two thirds red, then of course the evidence becomes more and more able to outweigh a previously held contrary opinion, whatever its nature.

Recently there have been developed powerful new methods of dealing with situations of this general sort, in which one wishes to draw all the justified inferences out of experimental evidence. Although Bayes' theorem cannot be applied unless one possesses or assumes prior opinions, it has been found that other procedures, associated with statistical theory rather than pure probability, are capable of drawing most useful conclusions.

What does a probability mean? What does it mean, for instance, to tell a patient: "If you decide to submit to this surgical operation, the probability that you will survive and be cured is .72"? Obviously this patient is going to make only one experiment, and it will either succeed or fail. What useful sense does the number .72 have for him?

The answer to this—and essentially to any question whatsoever that involves the interpretation of a probability—is: "If a large number of individuals just like you and just in your present circumstances were to submit to this operation, about 72 out of every 100 of them would survive and get well. The larger the number of individuals, the more likely it is that the ratio would be very close to 72 in each 100."

This answer may at first seem a little artificial and disappointing. It admittedly involves some entirely unrealizable conditions. A complicated intuitive process is required to translate the statement into a useful aid to the making of decisions. But experience does nevertheless show that it is a useful aid.



A theory may be called right or wrong according as it is or is not confirmed by actual experience. In this sense, can probability theory ever be proved right or wrong?

In a strict sense the answer is no. If you toss a coin you expect to get about half heads. But if you toss 100 times and get 75 heads instead of the expected 50, you have not disproved probability: probability theory can easily reckon the chance of getting 75 heads in 100 tosses. If that probability be written as 1/N, you would then expect that if you tossed 100 coins N times, in about one of those N times you would actually get 75 heads. So suppose you now toss 100 coins N times, and suppose that you get 75 heads not just one or two times, as you expect, but say 25 times! Is probability now disproved?

Again no. For the event that has now occurred, although amazingly rare, is still an event whose probability can be calculated, and while its probability is exceedingly small, it is not zero. Thus one goes on, again making a new ex-

THREE-CARD GAME, classically known as The Problem of Three Chests, illustrates deceptiveness of probability. One card is white on both sides; the second is white on one side and red on the other; the third is red on both sides (A). The dealer shuffles the cards in a hat (B), takes one out and places it flat on the table. The side showing is red (C). The dealer now says: "Obviously this is not the white-white card (D). It must be either red-white or red-red (E). I will bet even money that the other side is red." It is a poor bet for anyone else. Actually there are three possible cases (F). One is that the other side is white. The other two are that it is one or the other side of the red-red card. Thus the chance that the underside is red is 2 to 1.

periment which consists of many repetitions of the previous experiment. And even if miracles persist in occurring, these would be, from the point of view of probability, not impossible miracles.

Thus in a strict sense probability cannot be proved either right or wrong. But this is, as a matter of fact, a purely illusory difficulty. Although probability cannot be strictly proved either right or wrong, it can be proved useful. The facts of experience show that it works.

There are two different—or at least apparently different—types of problems to which probability theory applies. For the first type of problem probability theory is used not so much because we are convinced that we have to use it but because it is so very convenient. For the second type, probability theory seems to be even theoretically unavoidable. We shall see, however, that the distinction between the two cases, while of practical value, is really something of an illusion.

The first type has to do with situations which may be considered deterministic but which are so complex that the outcome is for all practical purposes unpredictable. In this kind of situation we realize that the final result has depended, often in a very sensitive way, on the interaction of a large number of causes. Many of these causes may be somewhat obscure in character, or otherwise impractical of detailed study, but it is at least thinkable that science could, if it were worth-while, analyze every cause in turn and thus arrive at a theory which could predict and explain what happens. When, in such circumstances, we say that the main final result "depends upon chance," we merely mean that, conveniently for us, the very complexity that makes a detailed analysis practically impossible assures an over-all behavior which is describable through the laws of probability.

Perhaps tossing a coin is again the simplest and most familiar illustration of this kind of case. There seems to be no essential mystery about why a coin lands heads or tails. The exact position of the coin above the table, the velocities of movement and spin given by the fingers, the resistance of the air, and so on—one can state what he needs to know in order to compute, by well-known dynamical laws, whether the coin will land heads or tails. But such a study would be very complicated, and would require very precise and extensive quantitative information.

There are many situations of this sort in serious everyday life, where we use probability theory not because it is clear that "chance" plays some obscure and mysterious role but primarily because the situation is so complicated, so intricately affected by so many small causes, that it is prohibitively inconvenient to attempt a detailed analysis. The experience of insurance companies, the occurrence of telephone calls and the resulting demands on telephone traffic and switching equipment, the sampling techniques used when one wishes to estimate the quality of many objects or the opinions of many individuals, the ordinary theory of errors of measurement, problems in epidemiology, the kinetic theory of gases—all these are practical instances in which

the causes are too numerous, too complicated, and/or too poorly understood to permit a complete deterministic theory. We therefore deal with these subjects through probability. But in all these cases we would say, with Poincaré, that chance "is only the measure of our ignorance."

The second type of probability problem at first sight seems very different. Most scientists now believe that some of the most elementary occurrences in nature are essentially and inescapably probabilistic. Thus in modern quantum theory, which forms the basis of our working knowledge of the atom, it seems to be not only impossible but essentially meaningless to attempt to compute just where a certain electron will be at a certain instant. All that one can do is reckon, as through the Schrödinger wave equation, the values of a probability position function. One cannot predict where the electron will be—one can only compute the probability that it will or will not be at a given place or places. And any attempt to frame an experiment that would resolve this probability vagueness, by showing just where the electron is, turns out to be a self-defeating experiment which destroys the conditions under which the original question can be asked.

It is only fair to remark that there remain scientists who do not accept the inevitable role of probability in atomic phenomena. The great example, of course, is Einstein, who has remarked in a characteristically appealing way that "I shall never believe that God plays dice with the world." But it is also fair to remark that Einstein, for all his great genius, is in a small minority on this point.

The problems that involve probability in this inescapable way are of the most fundamental kind. Quantum theory and statistical mechanics, which combine to furnish a large part of the basic theory of the physical universe, are essentially built upon probability. The gene-shuffling which controls inheritance in the living world is subject to probability laws. The inner character of the process of communication, which plays so great and so obvious a role in human life, has recently been found to be probabilistic in nature. The concept of the ever forward flow of time has been shown to depend upon entropy change, and thus to rest upon probability ideas. The whole theory of inference and evidence, and in fact of knowledge in general, goes back to probability.

We are now in a position to see that the two types of probability problems are, if we wish to be logically precise, not so different as we first supposed. Is it correct to think of the fall of a coin as being complicated but determinate, and the position of an electron as being essentially indeterminate? Obviously not. From a large-scale and practical point of view, one could doubtless deal quite successfully with coin-tossing on the basis of very careful actual measurements, plus all the analytical resources of dynamical theory. It remains true, however, that the coin is made of elementary particles whose positions and motions can be known, as science now views the matter, only in a probability sense. Thus we refine our original distinction between the

two types of cases by saying that the second type, whatever the scale involved, is essentially indeterminate, whereas the first type involves large-scale phenomena which may usefully be considered determinate, even though these large-scale phenomena depend ultimately on small-scale phenomena which are probabilistic.

Science deals (as in mathematics) with statements about theory which are logically accurate, but to which the concept of "truth" does not apply; it also deals (as in physics) with statements about nature which can never in a rigorous sense be known to be true, but can at best only be known to be highly probable. It is rather surprisingly the case that the only time man is ever really sure is not when he is dealing with science, but when he is dealing with matters of faith.

There is, moreover, some justification for saying that in science probability almost plays the role that faith does in other fields of human activity. For in a vast range of cases in which it is entirely impossible for science to answer the question "Is this statement true?" probability theory does furnish the basis for judgment as to how likely it is that the statement is true. It is probability which, in an important fraction of cases, enables man to resolve the paradoxical dilemma pointed out by Samuel Butler: "Life is the art of drawing sufficient conclusions from insufficient premises."

✦✦✦✦✦✦✦✦✦✦✦✦✦ MATHEMATICAL PROBABILITY

ERNEST KURNOW, GERALD J. GLASSER,
AND FREDERICK R. OTTMAN

Probability, chance, and likely are three words which are quite common in everyday conversation. We have all made, or at least heard, statements such as: the probability of a head when a coin is flipped is one half; the likelihood that there is life on Mars is low; the chances that the Yankees or the White Sox will win the pennant next year are very high; and the probability of holding four aces in a game of poker is less than the probability of holding a royal flush.

In which of these connections or in what other situations have you expressed similar ideas? What did you mean by them? It will be advantageous for you to give a few minutes of thought to these questions. Later you can check your intuitive notions against the more formal development given below. But for the present, your intuition should help you to appreciate some of our introductory remarks.

Our present aim is not to consider the term "probability" and its synonyms

Statistics for Business Decisions, *R. D. Irwin, Inc., 1959, 127–150.*

in their everyday context. Rather it is to give these ideas a technical, rigorous, and specific meaning. Thus, succeeding sections present some of the basic principles of *mathematical probability* as they have been laid down by mathematicians and statisticians over the past two or three centuries. This type of probability may not conform exactly to your intuitive feeling for the term as it is used in everyday conversation, but it should be somewhat similar. In any event, our short survey will involve, initially, the principles of the theory of probability as a branch of mathematics. Secondly, it will involve the applications of the results of mathematical probability in the field of statistics.

Incidentally, the difference between mathematics and statistics is borne out in the last two sentences. In short, the development of the theory of probability is mathematics, while one of the applications of this theory is statistics. The distinction may be clearer if we draw an analogy with geometry. This subject, as you may know, is developed mathematically; that is, its development hinges on mathematical principles: reasoning from axioms and postulates. The task of applying the results of this development to problems in the real world then falls to the engineer, the architect, the astronomer, or others in applied fields. In exactly the same way it is the statistician who may apply the mathematically derived results of probability theory. Nevertheless, it must be remembered that the foundation of statistics and statistical applications is the mathematical theory of probability.

As we have mentioned, mathematicians and statisticians like to give a more precise and definite meaning to probability than do most people in everyday conversation. What, then, is their definition of probability? This is an easier question to ask than to answer in an elementary textbook. We might hint at the difficulty by saying that there are many ideas on how probability should be defined. There is, however, one point of almost universal agreement. Almost everyone in the field of mathematics and statistics agrees that mathematical probability should be expressed quantitatively. Thus, their attention is centered upon a numerical measure of probability, or as it is sometimes called, a *probability measure.*

The setting for our problem is paralleled by a very familiar phenomenon, the weather. In everyday conversation, we often use the words "hot," "cold," "warm," "cool," "pleasant," etc. These have only a very general qualitative meaning. On the other hand, we are able to measure temperature on either a centigrade or a Fahrenheit thermometer. We are able to give temperature a quantitative meaning. This is precisely what we want to do with probability—to measure it on a scale. Thus, we shall interpret our probability measure as follows:

The probability of an outcome, A, is defined as the proportion of times A would occur in an infinite series of repeated trials of the same kind.

If we let A be some specified outcome in a random experiment, the symbol $P(A)$ will refer to the probability

that A will occur. Then our definition may be expressed in a formula as

$$P(A) = \frac{\text{Number of Times } A \text{ Occurs}}{\text{Number of Repeated Trials of the Same Kind}}$$

Thus, when we talk about the probability of an outcome, we mean the relative frequency with which it would occur if the experiment were repeated an infinite number of times. For example, we defined simple random sampling as a method of selection which gives every possible sample of size n the same chance of being chosen. What does this mean? It means that to select such a sample we must use a method which *if* repeated indefinitely would result in each combination appearing with the same relative frequency.

There are two aspects of the definition of probability that we must discuss: first, the notion of "infinite series" and "long run"; and, secondly, the idea of "repeated trials of the same kind."

The idea of "infinite series" and "long run" may be cleared up by a conventional illustration. In terms of our definition, what would we mean if we said that the probability of a coin falling heads when it is flipped is equal to .50? We mean that if this same process, flipping the coin, were continued indefinitely under the same conditions, we would expect 50 per cent of the outcomes to be heads.

We should also point out what the .50 does not mean. It does not necessarily imply that as we continue tossing the coin, we will come to a stage when the number of heads will always be equal to the number of tails. In fact, the difference between the number of heads and number of tails may very well get larger as we increase the number of flips. Thus, in 100 tosses of the coin we might obtain 60 heads and 40 tails, or an excess of 20 heads. After 1,000,000 tosses, we might have 500,500 heads and 499,500 tails, or an excess of 1,000 heads. The relative frequency of heads, however, has decreased from .60 to .5005—very close to our seemingly "good" estimate that the probability of a head is .50.

How do we know that the probability of a head is .50 when we flip a coin? The answer to this question is that we don't know—50 per cent is only a conjecture based on the apparent symmetry of the coin. We would know the true probability exactly only after observing an unlimited number of flips of the coin. This obviously is impossible. A little thought, in fact, will lead you to the conclusion that we can never determine any probability exactly. To do so would require an infinity of observations! What can be done so that we can ease ourselves out of this quandary? Fortunately, even though we can never know any probabilities exactly, there are various procedures for estimating them.

One method of estimating probabilities is on the basis of past experience. Two examples will illustrate this idea.

Example 1. Suppose that an office manager wishes to determine the probability of a paper being misfiled by his file clerks. He examines 1,000 papers filed during the past week and finds 30 misfiled papers. On the assumption that the process of filing or misfiling papers is a random process, an estimate of the probability of a given paper being misfiled, $P(A)$, is

$$P(A) = \frac{30}{1{,}000} = .03.$$

This value indicates that in the long run approximately .03 of the papers will be misfiled. Although the office manager cannot predict which individual papers will be misfiled on the basis of this information, the estimate does give him some insight into the efficiency of the filing process.

Example 2. In setting their rates, life insurance companies make use of mortality tables which give them the probability of a person dying during any one year of his life. These tables are based on past experience, gained by keeping records on ages at which persons die.

For example, one mortality table (there are several, but all give similar probabilities) lists the probability of a 21-year-old person living to 22 as approximately .992. Since he can only live or die, the probability of his dying before reaching 22 is approximately .008.

Here again the process is viewed as a random process. Probabilities such as these, of course, do not give probabilities strictly applicable to any one individual, who might be rather sick or very, very healthy. The probabilities refer to the population at large.

The second method of estimating probabilities is one based on the nature of the process we are observing. For example, when a coin is flipped, there are two possible outcomes, heads and tails; when a die is thrown, there are six possibilities, 1, 2, 3, 4, 5, and 6. Ordinarily, coins and dice are *symmetrical.* Unless we have reason to believe otherwise, it seems logical to *assume* that each possible outcome is *equally likely* or *equiprobable.* Thus, if a die is thrown, what is the probability a 4 shows? Using our equiprobable assumption, $P(4)$ equals 1/6 since there is only one way a 4 can occur and six different outcomes are possible.

As an illustration of a statistical application of the equiprobable assumption, reconsider the use of random number tables for selecting probability samples. Because of the nature of these tables and the ways in which they are derived, it is reasonable to assume that every digit, 0, 1, 2, . . . , 9 is equally likely in such a table. That is, we may assume that the probability of any digit occurring is 1/10. Likewise, we may assume that every pair of digits, 00, 01, 02, . . . , 99, is equally likely—that is, each has a probability of 1/100 of occurring.

The equiprobable rule for estimating probabilities must be used with care. For example, just because you can pass or fail your course in statistics, you wouldn't care to say that both outcomes are equiprobable.

We now turn to the second idea in our definition—the matter of repeated trials of the same kind. This requires some qualification because we have assumed that trials can be repeated. However, a man can only live or die once during his 21st year; in testing a given light bulb it will or won't burn out in its first 100 hours of use; etc. These experiments cannot be repeated. This presents a problem when we attempt to estimate a probability on the basis of past experience because if a trial cannot be repeated we wind up with very little experience.

The problem is usually resolved by pooling many trials of the *same kind.* To estimate probabilities of living and dying we observe not one individual

but many individuals of the same kind (although no two individuals are the same, we group those with the same age, occupation, race, etc.). When we flip a coin repeatedly, we assume that the flips are of the same kind (although they may not be exactly so; for example, the coin wears and/or your finger gets calloused).

No trial is exactly repeatable; no two trials are of exactly the same kind in any business situation. Thus, in determining the probability of a misfiled paper in Example 1, the trials might not have been of the same kind. The file clerks might have changed, the lighting system could have been improved, or a new supervisor might have been appointed. All these factors might result in trials not of the same kind. The best we can do is try to lump together those that are approximately of the same kind when we must estimate a probability. But even barring this remedy, our definition of probability still provides us with a theoretical bench mark. Remember, then, that $P(A)$ means the proportion of times A would occur *if* the experiment *were repeated* an unlimited number of times under the same conditions.

PROPERTIES OF A PROBABILITY MEASURE

Now, we turn to discuss three basic properties of a probability measure. These properties follow almost immediately from our definition of probability.

First of all, what are the numerical limits of a probability measure? What are its minimum and maximum values? The answer is that for any outcome the probability measure may have a value as small as 0 or as large as 1.

These extreme values of 0 and 1 have a convenient interpretation. Thus, if an outcome never occurs, no matter how many trials are attempted, the probability is considered 0. This is reasonable in view of our interpretation of probability; if an event never occurs then its relative frequency is 0. Furthermore, if one outcome always results from an actual experiment, the probability of this outcome may be considered as equal to 1. This, too, is consistent with the relative frequency definition of probability, as you easily can verify.

For example, on the basis of past experience we might estimate the probability of a baby being born with green hair as 0. On the other hand, we might also say that the probability that every person alive today will die at some time is 1.

A second property of a probability measure is that the sum of the probabilities of all possible outcomes is equal to 1. For example, suppose that we select at random one entry from an accounts payable ledger in order to check its accuracy. The two possible outcomes of this experiment are the selection of an incorrect entry and the selection of a correct entry. Therefore, the probability of an incorrect entry plus the probability of a correct entry must equal 1. Thus, if the probability of selecting a correct entry is .8, the probability of selecting an incorrect entry must be .2, since they are the only possible outcomes. Furthermore, suppose we were to select two entries from

the accounts payable ledger. Then 0, 1, or 2 of these may be incorrect. Whatever the probability of each of these outcomes, their sum must be 1.

The third and last property of a probability measure refers to the likelihood that one or the other of two possible outcomes occurs. We will consider this property only as it applies to outcomes that are *mutually exclusive.* Mutually exclusive outcomes are *those which cannot occur at the same time.* Some illustrations should fix the idea in your mind.

Example 3. A personnel administrator must decide on one of several applicants for a position with his company. His alternatives, which include hiring J. Smith, H. Jones, M. Jackson, etc., are mutually exclusive since he is to hire only one of the applicants.

A worker may be absent exactly 0, 1, 2, 3, 4, or 5 days in a given week. All of these possibilities are mutually exclusive. Thus, for example, if he is absent exactly twice, he cannot be absent exactly once.

A simple random sample of n students is to be selected and their average age calculated. Among the possible outcomes of this random experiment are $\bar{x} = 22.1$ years and $\bar{x} = 20.5$ years—two outcomes which are mutually exclusive. If the sample mean is 22.1 years, it cannot be 20.5 years.

A simple random sample of 2 machines from a group of 6 machines (A, B, C, D, E, and F) is selected. If the sample AB occurs, AC cannot occur—nor can AD, AE, etc. All of these outcomes are mutually exclusive.

The rule expressing the probability of mutually exclusive outcomes may be stated as follows: *The probability of either of two mutually exclusive outcomes occurring is equal to the sum of their individual probabilities.* If we write $P(A \text{ or } B)$ for "the probability of either A or B," we can summarize this rule as

$$P(A \text{ or } B) = P(A) + P(B).$$

This rule is sometimes called the *addition rule.* We must remember that this rule is valid only if the two outcomes are mutually exclusive.

This rule may be extended to three or more mutually exclusive outcomes as well. Thus, if A, B, C, etc., are all mutually exclusive:

$$P(A \text{ or } B \text{ or } C \text{ or } \ldots) = P(A) + P(B) + P(C) + \ldots.$$

Example 4. Suppose that the probabilities of exactly 0, 1, 2, 3, 4, etc., machines being out of order at the same time in a plant have been estimated on the basis of past experience. Let $P(0)$ be the probability of 0 machines being out of order, $P(1)$ the probability of 1 machine, etc. Then the information can be summarized as follows:

$$\begin{aligned} P(0) &= .449 \\ P(1) &= .360 \\ P(2) &= .144 \\ P(3) &= .038 \\ P(4) &= .008 \\ P(5 \text{ or more}) &= .001. \end{aligned}$$

What is the probability of exactly 0 or exactly 1 machine being out of order at the same time? Since 0 and 1 are mutually exclusive outcomes, the addition formula can be applied:

$$\begin{aligned} P(0 \text{ or } 1) &= P(0) + P(1) \\ &= .499 + .360 \\ &= .809. \end{aligned}$$

What is the probability of 2 or less machines being out of order? Similar reasoning shows that

$$P(0 \text{ or } 1 \text{ or } 2) = P(0) + P(1) + P(2)$$
$$= .449 + .360 + .144$$
$$= .953.$$

You will recall that the addition rule is *not* applicable when the outcomes are not mutually exclusive. For example, if you were to select a student at random from your statistics class, the probability of selecting a junior or a student over 21 years of age does not refer to mutually exclusive outcomes. A student may be both a junior and over 21 years at the same time; one outcome does not exclude the other. Other simple rules have been developed to cover the calculations of probabilities for nonmutually exclusive outcomes, but we shall not consider them. . . .

REPEATED TRIALS— CONDITIONAL PROBABILITY

Up to this point we have dealt with probabilities of outcomes in a single random experiment or trial. For example, we spoke of the probability of selecting a correct entry in an accounts payable ledger, and of the probability of 1 *or* 2 machines being out of order at one time. In this section, we shall determine how to compute the probability of a given sequence of outcomes in repeated experiments or trials.

Example 5. A coin is tossed 3 successive times—these represent 3 repeated trials. One question dealing with these trials that might be of interest is: "What is the probability of the sequence head-head-head?"

An employee is to report to work 5 days in a given week—this situation presents 5 repeated trials. Each trial may result in the worker being absent or present. The type of question we will attempt to answer in this section is exemplified by "What is the probability of the sequence absent-absent-present-present-present?"

Three machines are observed for a week to determine whether they require maintenance work or not. Each machine represents 1 trial—the 3 machines represent 3 trials.

In developing rules for computing probabilities of a given sequence of outcomes from a series of repeated trials, there are two basic ideas to keep in mind. First of all, we desire rules for expressing the probability of a sequence of outcomes in terms of the probabilities of individual outcomes. This situation is similar to the one we met and solved . . . [under "Properties of a Probability Measure"] where $P(A \text{ or } B)$ was expressed in terms of $P(A)$ and $P(B)$—as their sum. Thus, to compute probabilities of sequences we will have to know individual probabilities in order to apply the rules to be developed.

Secondly, it is useful to distinguish between two types of repeated trials and to develop two different rules for these types. Thus, repeated trials may be *dependent* or *independent*, depending upon whether any one trial influences the others or not. For example, two successive flips of a coin should not influence each other—hence, they are independent. On the other hand, suppose that we were to draw two cards from a deck. The outcome of the first draw would influence the outcome of the second (e.g., if the ace of clubs were drawn first it could not be drawn second). Thus, these trials would be dependent.

In this section, we shall develop a rule for computing the probability of a sequence of outcomes when the trials are dependent; later we shall consider independent trials. Thus, the rule for calculating the probability of a specified sequence of two outcomes in two dependent trials states: *The probability of a sequence of two outcomes is equal to the probability of the first outcome multiplied by the probability of the second outcome given that the first outcome has occurred.* You will observe that the probability of the second outcome is conditional upon the occurrence of the first outcome. Hence, it is called a conditional probability. The rule can be stated symbolically as follows:

$$P(A \text{ and } B) = P(A) \times P(B|A).$$

The vertical line in $(B|A)$ stands for the word "given." Let us look at an illustration of the rule.

Example 6. Two forms of a general achievement test are to be selected at random from a group of 6. These tests are to be administered to entering college freshmen. The forms are numbered 1 through 6 inclusive. What is the probability that forms 2 and 4 will be selected *in that order?* When we select the first form of the test, we have 6 tests to draw from. If we use random numbers, the probability of form 2 being drawn first is 1/6. However, if form 2 is selected, the next test must be drawn from the remaining 5. The probability of drawing form 4 is 1/5. The drawing of form 4 as the second selection is conditional upon the drawing of form 2 first. The probability that forms 2 and 4 will be drawn in that order is

$$\begin{aligned} P(2 \text{ and } 4) &= P(2) \times P(4|2) \\ &= \frac{1}{6} \times \frac{1}{5} \\ &= \frac{1}{30}. \end{aligned}$$

This answer may be verified by enumerating the set of all possible sequences of two tests as follows:

1,2	2,1	3,1	4,1	5,1	6,1
1,3	2,3	3,2	4,2	5,2	6,2
1,4	2,4	3,4	4,3	5,3	6,3
1,5	2,5	3,5	4,5	5,4	6,4
1,6	2,6	3,6	4,6	5,6	6,5

When a random method of selection is used, each of these 30 different sequences is equally likely. Form 2 followed by form 4 represents only 1 of the 30 possible outcomes, and, hence, its probability of occurrence is 1/30—the same value as that which was obtained by the application of the conditional probability rule.

The conditional probability rule can be generalized for more than two outcomes as follows:

$$P(A \text{ and } B \text{ and } C \text{ and } \ldots) = P(A) \times P(B|A) \times P(C|A \text{ and } B) \times \ldots,$$

where $P(C|A \text{ and } B)$ refers to the probability of C given that A and B have occurred, etc. Thus the probability of a sequence of outcomes in repeated random trials is equal to the product of the probabilities conditional upon the occurrence of the preceding outcomes in the sequence. For example, reconsider the situation described in Example 6 and assume 3 test forms are to be selected. What is the probability that forms 2, 4, and 5 would be selected *in that order?* The probability of select-

ing form 2 first is 1/6; and the probability of selecting form 4 second, given that form 2 has been selected first, is 1/5. In addition, the probability of selecting form 5 third, given that forms 2 and 4 have already been drawn, is 1/4. Hence, the probability of forms 2, 4, and 5 being chosen in that order is

$$P(2 \text{ and } 4 \text{ and } 5) = \frac{1}{6} \times \frac{1}{5} \times \frac{1}{4} = \frac{1}{120}.$$

The conditional probability rule for two outcomes and its extension for three or more outcomes are quite useful in the development of statistical theory. . . .

INDEPENDENT REPEATED TRIALS

When repeated trials are independent rather than dependent, a special case of the conditional probability rule may be applied. This special case is frequently encountered in statistics, but first may be illustrated by reference to a game of chance. For example, what is the probability of obtaining two heads in two tosses of an unbiased coin? Applying our conditional probability rule

$$P(H \text{ and } H) = P(H) \times P(H|H).$$

That is, the probability of a sequence of two heads is equal to the probability of the first head multiplied by the probability of a second head given that the first toss is a head. Thus,

$$P(H \text{ and } H) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

You will observe that since the coin is unbiased, the probability of a second head is not influenced by the occurrence of the first head. The probability is 1/2 in both cases. In other words, the conditional probability $P(H|H)$ is the same as the unconditional probability, $P(H)$. This always is the case when the outcomes under consideration result from independent trials.

We now can restate the conditional probability rule for the special case. When trials are independent,

$$P(A \text{ and } B) = P(A) \times P(B).$$

The rule can be extended to more than two trials as follows:

$$P(A \text{ and } B \text{ and } C \text{ and } \ldots) = P(A) \times P(B) \times P(C) \times \ldots .$$

We are especially interested in the rule for independent trials when each trial can result in one of two mutually exclusive outcomes. For example, a coin can be a head or a tail; a television cabinet can be defective or nondefective; a person can be male or female; a family's income can be less than $10,000 a year, or $10,000 or more a year. In our discussion of repeated independent trials we will concern ourselves only with such outcomes.

The rule for independent trials is applicable in answering questions such as the following: What is the probability that in a simple random sample of 100 entries from an accounts payable ledger there will be 2 incorrect entries and 98 correct entries? Suppose that we are concerned with the system of recording ledger entries. Hence, we are dealing with an analytical study. Our population is, therefore, infinite—consisting of the ledger entries past, pres-

ent, and future made under similar conditions. The probability of an entry being selected as the second unit in a sample is not conditional upon the selection of the first entry. When 1 entry is chosen, there still remain an unlimited number of entries.

Suppose that we reconsider an earlier illustration to indicate the use of the rule for independent trials in the evaluation of the risks of relying on a decision rule.

Example 7. A firm which has a policy of attempting to fill orders within a week's time has reasons to suspect that its shipping process has slowed down. In the past it has been determined that only 5 per cent of its orders were delayed more than a week—and this is regarded as satisfactory. Now, however, several customer complaints hint at the possibility that π, the proportion of delayed orders, has increased from its normal and satisfactory level of .05.

Note that this situation calls for an analytical study. It deals with a (shipping) process and, therefore, presents an infinite population.

Suppose that the firm plans to select a sample of 3 orders as a basis for deciding on whether or not to attempt to improve the shipping process. How will they use the information obtained from this sample? If none of the 3 shipments is delayed, the shipping process will not be checked; if 1 or more are delayed, the process will be investigated with an eye towards revising it. This is their statistical decision rule—but what is the risk of applying it? Actually there are many risks.

One is the risk of investigating the process because the sample shows 1 or more delayed orders, even though π has remained at its normal and satisfactory level of .05. If we assume that the 3 orders selected for the sample are filled independently, this risk may be computed by the use of the rule for independent trials. Thus, if $\pi = .05$, for a single order selected at random,

$$P(\text{Delayed Order}) = .05$$
$$P(\text{Undelayed Order}) = .95.$$

Then the chance of not investigating the process is the chance of selecting 3 successive undelayed orders, namely,

$$\begin{aligned} &P(\text{Undelayed and Undelayed and Undelayed}) \\ &= P(\text{Undelayed}) \times P(\text{Undelayed}) \times P(\text{Undelayed}) \\ &= .95 \times .95 \times .95 \\ &= .857. \end{aligned}$$

It follows that the probability of investigating if $\pi = .05$ is $1 - .857 = .143$. This is the risk of checking a normal and satisfactory process.

Another risk is that the sample will lead to not investigating the shipping process when π actually has increased significantly. For example, suppose $\pi = .25$. Then,

$$P(\text{Delayed Order}) = .25$$
$$P(\text{Undelayed Order}) = .75$$

and the chance of not investigating the process is

$$.75 \times .75 \times .75 = .422,$$

while the chance of investigating is $1 - .422 = .578$.

Similar calculations indicate some of the other probabilities of investigating and of not investigating the process are as follows:

π	P(*Investigating*)	P(*Not Investigating*)
.00	.000	1.000
.05	.143	.857
.10	.271	.729
.15	.386	.614
.20	.488	.512
.25	.578	.422

Check one or two of these values. Also, notice that these probabilities are all characteristics of the decision rule under consideration. To evaluate that rule you should consider these chances or risks. Do they seem economical or not? Why or why not?

The last illustration affords an opportunity for us to make note of an important thought. You can see that the probability of investigating the process increases as π increases. Generally speaking, this means that the more the shipping process has deteriorated, the better our chances of taking the right action—of checking the process. Thus, for $\pi = .10$, this probability is .271, while for $\pi = .25$, it is .578. For less and less desirable states of the world the probability of taking the right action gets higher and higher, and the risk of making an incorrect decision gets lower and lower. This is true despite the fact that the sample size is quite small. What would a larger sample accomplish? It would mean smaller risks for the different values of π.

THE BINOMIAL FORMULA

In the preceding section we used the rule for independent trials to evaluate the risks in relying on sample information from an infinite population. While the method of calculation we used to compute these probabilities is relatively simple for small samples, a more generalized procedure would facilitate the determination of probabilities of outcomes for any size of sample.

Such a generalized procedure is afforded by mathematical probability. It takes the form of a formula called the *binomial formula.* This formula simply represents an extension of the rule for independent trials. It also is related to the binomial theorem or binomial expansion, a relatively simple mathematical relationship of importance in algebra. Perhaps you remember studying the binomial; if so, you will recognize the relationship between it and what we are studying here. If not, don't worry about it—just remember that the formula we are studying is called the binomial.

Let us indicate the derivation of this formula. Suppose that we have n independent trials, such as n selections from an infinite population, and that each of the n trials can result in only one of two outcomes. To keep a concrete example in mind, envision a production process producing n parts each of which may be either defective (D) or nondefective (N) and assume these parts are produced independently.

What is the probability of a particular sequence? If we let

$$P(\text{One Defective Part}) = \pi$$
$$P(\text{One Nondefective Part}) = 1 - \pi,$$

then by the rule for independent trials, the probability of the sequence D, N, D is

$$\pi \times (1 - \pi) \times \pi,$$

or, in short,

$$\pi^2(1 - \pi).$$

Similarly, the probability of the sequence N, D, D is

$$(1 - \pi) \times \pi \times \pi = \pi^2(1 - \pi),$$

while that of the sequence D, D, N is

$$\pi \times \pi \times (1 - \pi) = \pi^2(1 - \pi).$$

You should note that all three of these sequences refer to 1 nondefective and 2 defectives—although they refer to their occurrence in different *orders*. It is no surprise, then, that the probability of each is $\pi^2(1-\pi)$. Now, it is important to recognize that the outcome D occurs twice in each sequence, and, hence, the exponent on π, the probability of D occurring, is 2. Since N occurs once the exponent on $(1-\pi)$, the probability of N occurring, is 1.

Suppose, now, that we are not interested in the order of occurrence and that we simply need to know the probability of 1 nondefective and 2 defectives in three trials. This is the probability of observing *N D D or D N D or DDN* and equals

$$\pi^2(1-\pi)+\pi^2(1-\pi)+\pi^2(1-\pi)$$
$$=3\pi^2(1-\pi)$$

because they are mutually exclusive possibilities.

Each of the terms, 3, π^2, and $(1-\pi)$ has an interpretation. We have already noted that the exponents on π and $(1-\pi)$ represent the frequency of D and N in a sequence. A coefficient, such as the 3, indicates the number of sequences in which, say, 1 nondefective and 2 defectives can occur.

Thus, if we were interested in the probability of 2 defectives and 2 nondefectives, we know that we must have $\pi^2(1-\pi)^2$ times some coefficient. What does this coefficient equal? There are six orders of two D's and two N's, namely,

D D N N	*N D D N*
D N D N	*N D N D*
D N N D	*N N D D*

and hence the coefficient is 6. Thus, we may write

$$P(\text{2 Defectives and 2 Nondefectives})$$
$$=6\pi^2(1-\pi)^2.$$

The binomial is a generalized formula for computing the probability of a given series of outcomes *in any order*. Thus, consider n trials in which

r are defective, and
$n-r$ are nondefective.

The probability of any such sequence is

$$\pi^r(1-\pi)^{n-r}$$

since the exponent on π refers to the number of outcomes that are defective and the exponent on $(1-\pi)$ refers to the number of outcomes that are nondefective.

Furthermore, the probability of r defectives and $n-r$ nondefectives *in any order* is $\pi^r(1-\pi)^{n-r}$ multiplied by the number of sequences in which r defective and $n-r$ nondefective can occur. This number of sequences may be expressed in terms of factorials as

$$\frac{n!}{r!\,(n-r)!}.$$

. . . With $n=4$ and $r=2$, we find

$$\frac{4!}{2!\,2!}=\frac{4\times3\times2\times1}{2\times1\times2\times1}=6,$$

a result which we also found earlier by direct enumeration.

Thus, if we write $P\left(\frac{r}{n}\right)$ for the probability of r outcomes of a specified type (such as defectives) in n trials,

$$P\left(\frac{r}{n}\right) = \frac{n!}{r!\,(n-r)!}\,\pi^r\,(1-\pi)^{n-r}.$$

This is the binomial formula. Remember that

n = number of trials,
r = number of specified outcomes,
$n - r$ = number of other outcomes,
π = probability of a specified outcome in a single trial.

The binomial, as noted, is a general formula for computing probabilities for independent trials, and as such is useful in statistics for computing risks of relying on sample information when the sample observations are independently derived attributes.

Example 8. Each hour a production manager selects a sample of 5 parts produced by a machine. If more than one of the parts is defective, he stops the process and resets the machine; if none or one are defective, he permits it to continue as is. What is the risk of stopping the process when π is only .02 (a quality level the firm considers satisfactory)? The binomial provides the answer. The probability of not stopping the process is the probability of finding 0 *or* 1 defectives in 5 trials. Hence, we should let

$$n = 5,\ r = 0,\ \pi = .02,$$

thus finding, for the probability of 0 defectives,

$$P\left(\frac{0}{5}\right) = \frac{5!}{0!\,5!}\,(.02)^0\,(.98)^5 = (.98)^5 = .904.$$

(Remember that $0! = 1$ and $(.02)^0 = 1$.)

Similarly, for the probability of 1 defective,

$$P\left(\frac{1}{5}\right) = \frac{5!}{1!\,4!}\,(.02)^1\,(.98)^4 = 5(.02)\,(.98)^4 = .092.$$

These two results may be combined to show the probability of 0 *or* 1 defective (which is the probability of not stopping the process):

$$.904 + .092 = .996.$$

Thus, the risk of stopping the process if $\pi = .02$ is only $1 - .996 = .004$. Does this mean that the production manager's rule is a good one? Why or why not? . . .

EXPECTED VALUES

. . . We have considered ways of measuring and computing probabilities of various outcomes in a random experiment or a series of random experiments. In this particular section we shall deal with an allied topic—*expected value.* This is the average value that would occur if the random experiment were to be repeated indefinitely.

For example, suppose a die is thrown and the face value observed. The average value observed after many throws is what is meant by the expected value. What is this expected value? Assuming each side equiprobable, we know that each face value—1, 2, 3, 4, 5, and 6—is equally likely; that is, the probability of each is 1/6. By assumption, then, in the long run each of these values would occur 1/6 of the time. The average value is thus,

$$\frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6) = \frac{21}{6} = 3.5,$$

and the 3.5 is called the expected value of the random variable under consideration.

Generally speaking, the expected value of any random variable may be

computed by (*a*) multiplying each possible value the variable may assume by its probability of occurrence, and (*b*) summing these products. This is the method used in the preceding paragraph—each value (1, 2, 3, 4, 5, and 6) was multiplied by its probability of occurrence (1/6), and these results were added together.

Example 9. Suppose that on the basis of his past experience, a newsdealer estimates the probabilities of selling 7, 8, 9, or 10 magazines per hour as follows:

$$\begin{aligned} P(7) &= .2 \\ P(8) &= .4 \\ P(9) &= .3 \\ P(10) &= .1. \end{aligned}$$

The expected number of magazines is then

$$7(.2) + 8(.4) + 9(.3) + 10(.1) = 8.3.$$

This is an average value—an average that might be expected in the long run.

Example 10. A manufacturer receives shipments of a special part in lots of 50. He uses the following decision rule to decide whether or not to inspect the entire lot:

> Select a simple random sample of 5 parts. If none of the 5 selected parts is defective, do *not* inspect the remaining 45 parts. If 1 or more are defective, inspect the remaining 45 parts.

The manufacturer desires to know the expected amount of inspection in order to estimate the cost of inspection under this decision rule.

The total amount of inspection for any one lot depends upon the action taken. Note the tabulation that follows:

Action	*Number of Items Inspected*		*Total Number of Parts Inspected*
	Sample	*Remainder of Lot*	
No further inspection	5	0	5
Full inspection	5	45	50

For any lot, however, the total number of parts inspected is a random variable since it results from a random experiment of testing a sample of 5 parts. To find the expected amount of inspection it is, therefore, necessary to determine the probability of each action under a given assumption about π.

Let us assume that a shipment contains 5 defectives—i.e., π(the proportion of defectives) is 5/50, or .10. Then under the decision rule

$$\begin{aligned} P(\text{No Further Inspection}) &= P(0 \text{ Defectives in a Sample of } 5) \\ &= \frac{45}{50} \times \frac{44}{49} \times \frac{43}{48} \times \frac{42}{47} \times \frac{41}{46} \\ &= .577. \end{aligned}$$

It follows, therefore, that for any lot with $\pi = .10$, the probability of fully inspecting a lot is: $1 - .577 = .423$.

We can compute the expected amount of inspection:

$$\begin{aligned} E(\text{Amount of Inspection}) &= (5 \times .577) + (50 \times .423) \\ &= 24.0. \end{aligned}$$

This means that if many lots with $\pi = .10$ are received, on the average 24.0 parts will be inspected.

However, you must remember that the expected amount of inspection depends on the probabilities of no inspection and of full inspection—and, hence, on the proportion of defectives in a lot. We have found the answer for $\pi = .10$, and similar calculations would enable us to determine the expected amount of inspection for other values of π. (For example, suppose $\pi = .00$. Then the probability of fully inspecting a given lot is 0. Why? Therefore, the expected amount of inspection is 5. Why?)

An expected value is a special type of arithmetic mean. That is, it is an average value. It is special only in the sense that it is the average value of a random variable—the average of numbers which are the result of a chance experiment.

Thus, it must be stressed that the term "expected value" does not refer to what we expect to occur on any one trial—in any single random experiment. An expected value is an average of what would happen only if such an experiment were repeated over and over again (an unlimited number of times). Thus, to refer back to the fact that the expected value of the outcome of throwing a die is 3.5, we can see it is impossible to get a 3.5 on a single toss. Certainly this is not what we should "expect" on any one throw. The 3.5 simply indicates the average value that we would expect to occur in the long run.

It is customary in mathematics and statistics to use the symbol E to represent the term "expected." Thus, if we let x refer to the face value that occurs when a die is thrown, $E(x)$ refers to the "expected value" of this experiment and $E(x) = 3.5$.

The concept of expected value is useful in many connections in probability theory and in its fields of application. In statistics, it is used primarily in connection with *evaluating sampling procedures* and, hence, in determining sound statistical decision procedures.

✦✦✦✦✦✦✦✦✦✦✦✦✦ CHANCE AND STATISTICS

HORACE C. LEVINSON

. . . Crude facts are the raw materials of statistics; from them is evolved the finished product, usually a table of chances. A crude fact is a simple statement of a past event, such as "John Smith died yesterday at the age of forty." From large numbers of such crude facts the insurance actuaries, for example, have constructed mortality tables, which serve as a basis for computing the amount of your life insurance premium. This basic table gives the chance that a person of such and such an age will

The Science of Chance, *1963, Dover Publications, 218–242.*

survive at least so many years. It tells us, among other things, that of one hundred babies born now, about seventy will be alive forty years from today. In other words, the chance that a child born now will reach the age of forty is approximately 7 in 10.

The choice of the crude facts that are admitted into the statistical table determines the scope and meaning of the resulting mortality table. If we compare the English, French, and American mortality tables, for example, we shall find no two of them alike. Each reflects the enormously complex factors, some of them peculiar to the country in question, and all of them beyond the powers of direct analysis, which determine the span of life.

If we include in the table only individuals who have been medically examined, and whose general health is determined to be above a specified level corresponding to the actual procedure in writing life insurance, the table will be significantly modified. If we admit into one statistical table only data concerning males, and into another only data concerning females, we shall find that the resulting mortality tables have important differences. If we construct mortality tables according to the occupations or professions of those whose vital statistics are included, we shall find, similarly, a number of distinct tables. If we adopted one of a large number of other classifications, there would result still other mortality tables. Which classifications are used in practice depends on the purpose for which the mortality tables are intended. The important thing is that the selection be made with this purpose in view. The corresponding thing is true not only in actuarial statistics, which is concerned with all forms of insurance, but in other statistical applications as well, whether the subject be biology or business.

Statistical tables are, then, the fruits of experience. They lead to the determination of probabilities which would otherwise be hopelessly inaccessible. . . . in many games of chance it is possible to compute the probabilities of the game entirely independently of all experience. After this has been done, these probabilities can be compared with the results of experience. . . . in more complicated matters experience, in the form of statistical tables, is the only possible road to a knowledge of the chances involved. In computing mortality tables, for instance, there can be no question of finding a priori probabilities. The causes that determine length of life are so complicated that we cannot even imagine what an analysis into "equally likely cases" would mean, although we see very clearly what they mean in throwing dice or dealing poker hands. In these complex situations we substitute *statistical probabilities*, determined by experience.

Not only are statistical tables the fruits of experience, but they are the fruits of large numbers of experiences. These experiences may consist of a great many repetitions of some event, as when one person throws dice a large number of times, or they may consist of the combined dice throwing of a large number of persons. In life insurance tables, the experience is necessarily of the second kind; for no one

can die repeatedly, no matter how willing he is to sacrifice himself for the benefit of science, or of the insurance companies. In games of chance we have been able to predict with confidence how a player will come out *in the long run,* or how a *very large number* of players will come out in a single trial. Similarly in statistics, if our tables are based on a very large number of experiences, we can, in many cases, predict with confidence what will happen in the long run, whether we are concerned with an individual repeating an experience, or with a group of individuals.

This, perhaps, sounds too generalized, but its actual application to statistical problems is not difficult. Suppose that we have before us a table giving the heights of one hundred thousand American males, selected at random, and that we are interested only in learning something about the average height of Americans. It is simple enough, although a trifle tedious, to add all the heights together, and divide by their number, which is 100,000. This gives us the average height of the males in our sample. What have we learned by doing this? Can we assert that the average height of adult American males is very close to 5 feet 8 inches, if that is the result of averaging the sample? Before being able to make such an assertion with confidence, we should have to satisfy ourselves on several points.

We have been informed that our table of heights contains individuals "selected at random." If we are conscientious statisticians, or critics of statistics, we shall never pass this by without somehow assuring ourselves that the phrase "selected at random" means approximately the same thing to the person or persons who collected the facts for the table, as it does to us. We are faced with the question . . . of determining whether a given sample is adequate to represent the entire population. Above all we should like to know on precisely what basis one man out of every four hundred was selected, and the other three hundred and ninety-nine omitted. With this information at hand, we shall soon be able to decide whether our idea of random selection corresponds to that used in making the table of heights. In the absence of precise information on this point, there are many questions to which we should require answers before being willing to attach weight to our conclusions: Does the sample correspond in geographical distribution to that of the population of the country? Does it include a representative distribution of ages? How many men of foreign birth are included? How many Negroes? Does the table show a preference for any particular class, social or occupational? On the other hand, if we knew that the selections were made by drawing names from boxes containing all the names, or in some equivalent way, we should feel satisfied that we are dealing with a fair sample.

Next, we ask whether a random sample of one hundred thousand is large enough for our purpose. We are attempting to reach a conclusion as to the adult male population of the United States, amounting to forty-odd millions of individuals, by using a sample containing measurements of only one hun-

dred thousand, or 0.25 per cent of the whole. Before this question can be answered at all, it is necessary to know the accuracy that is required of us. If it is required to know merely whether the average height of an American falls between 5 feet and 5 feet 6 inches, or between 5 feet 6 inches and 6 feet, it is clear enough that a small sample is all that is needed. If it is required to determine this average height closely enough to feel confident that we know it to the nearest inch, a much larger sample is necessary, and so on.

Once given the degree of accuracy that is required of us, there remains the problem of determining whether or not the sample of 100,000 measurements is large enough to furnish it. This problem is typical of a whole class of statistical questions, to answer which a part of the mathematical theory of statistics has been built. We are not concerned here with the precise mathematical methods used to solve such problems; they belong to the professional kit of the statistician and are set forth in technical treatises. We are concerned rather to catch the spirit of the methods, to grasp the character and significance of the conclusions reached, and above all to understand the precautions that must be taken before accepting these conclusions.

We have before us the list of 100,000 measurements of heights, and we have also the average of these heights. In amateur statistics, the study of an experience table usually begins and ends with taking the average. While it is true that the ordinary arithmetical average is an important first step, it is but one of several essential characteristics of a table, and it is seldom true that the average alone is of value.

It is by considering also features of the table other than the simple arithmetical average that the statistician reaches his conclusions as to the adequacy of the sample of 100,000 heights. His next step is to compute a quantity called the *standard deviation,* which measures the extent to which the heights are scattered about their average, in other words their *dispersion.* The amount of dispersion is evidently of major importance in the interpretation of the sample. We shall not stop here, however, to explain in detail how the standard deviation is computed; this will be done later in connection with a simple example, where the process can be easily illustrated. We shall merely remark that the standard deviation is itself an average.

In practice, when the statistician is confronted by a mass of data, such as the 100,000 heights, before making any computations whatever he arranges them in a frequency table. To do this he first chooses a limit of accuracy in listing the data, which he calls the *class interval.* Let us assume that in this case he selects half an inch. Then he enters opposite the height 68 inches, for example, the number of heights in the original data that lie between 67.75 inches and 68.25 inches. By making this table he shortens considerably the labor of computing the average; for he can first make a fairly accurate guess as to its value and then find out, by a mathematical process, how far off this guess is.

After entering each of the 100,000 heights in its appropriate place, he has

before him what is known as a *frequency distribution.* It tells at a glance how many men out of one hundred thousand in the sample are within 1/4 inch, in either direction, of any one particular height, say 6 feet. It also tells at a glance the most popular or "stylish" height, called the *mode,* opposite to which the greatest number of individuals are entered.

Suppose, for example, that opposite 6 feet, or 72 inches, we find the figure 2,000. This means that 1/50, or 2 per cent, of the individuals in the sample are 6 feet tall, to the nearest 1/2 inch. Assuming that the sample adequately represents the entire population, we can express this result as follows: The probability that an American male *selected at random* is 6 feet tall to the nearest 1/2 inch is 1 in 50. For each possible height other than 6 feet we can discover the corresponding probability in the same simple way. Thus our statistical table has led us to a table of chances.

What we wish to emphasize here is that the statistician's conclusion, when he reaches it, is in terms of chance or probability. He tells us the probability that the average value of 5 feet 8 inches is correct to within a certain amount, say 1/10 of an inch. He may say, "The betting odds are even that the true value of the average for the entire population is within 1/10 of an inch of that of your sample. It follows that the betting odds are 4½ to 1 that it is within 1/5 of an inch, and 100,000,000 to 1 that it is within 9/10 of an inch." This is the sort of answer that we might have expected in the beginning, and in fact, the sort that should satisfy us the most.

When statistics answers a question for you, always look for a tag of some sort carrying a reference to chance. Its absence is a clear danger signal.

In order to see as clearly as possible the intimate relation that exists between statistics and the theory of chance, let us examine a situation of an entirely different sort, where we shall obtain a glimpse of a statistical law in action. When a rifle is fired at a target, the shots group themselves in a pattern that is more or less spread out; no matter how good a shot is behind the gun, there is always a certain amount of this spreading, or *dispersion.* After a few shots have been fired, the marks on the target are usually quite irregularly placed; there is no semblance of order.

If, however, many rounds are fired, there appears a definitely ordered pattern. There is a center about which the shots are densely packed in all directions; farther from it there are less and less shots, until finally a circle is reached outside of which there are none at all. (We are assuming that the rifleman is a good enough shot to hit the target every time, or, what comes to the same thing, that the target is large enough for him to hit.) This center may or may not coincide with the bull's-eye. When it does not, there is a *systematic* error of some sort, such, for example, as an error in the sights. On the other hand, the errors that cause the spreading about the center of the pattern are called *accidental;* they are due to one or more of a large

number of causes, each small in itself: unsteadiness of hand, eccentricities of vision, variations of charge, irregular currents of air, and many others. The important thing about these accidental errors is their indifference to direction. Each one is as likely to be above as below the center of the pattern, to the right as to the left of it.

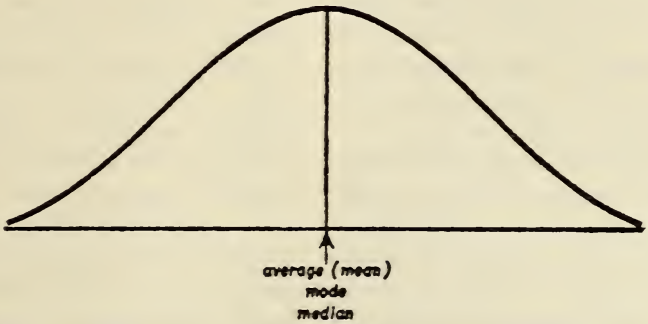


FIGURE 1

NORMAL FREQUENCY DISTRIBUTION (GAUSS'S LAW OF ERRORS)

In this illustration, which is typical of one class of those to which the statistical method can be applied, we see order emerging out of chaos. The more complete the underlying confusion, the more perfect the resulting order. Moreover, the ordering process is a progressive one; as more and more rounds are fired, it becomes more and more pronounced.

Precisely what is meant when it is said that the shot pattern is in accord with statistical law? Does it mean that the grouping of the shots about the center can be predicted in advance of the firing? Clearly not, for, apart from the systematic errors, nothing would then distinguish the expert shot from the man who has never had a rifle in his hands. Let us leave aside the systematic errors, which are utterly unpredictable and which can, for our purposes, be eliminated. Then the answer to our question is as follows: By use of the fact, mentioned above, that the accidental errors are indifferent to direction, and of one or two additional assumptions of a technical nature, it can be shown that the grouping of the shots is described by a statistical law known as Gauss's Law of Errors.

The mathematics involved in arriving at this result is beyond the scope of this book. It is important to understand, though, that such a law exists.

Gauss's law expresses the number of errors of each different size that will occur in the long run. It says that the smaller the error, the more frequently it will occur (the greater its probability, in other words) and that the probability of larger errors is progressively smaller. The curve drawn in Figure 1 above is an illustration of Gauss's law. Every law of this nature contains, in

addition to one or more variable quantities, the familiar x, y, and so on, of algebra, a few undetermined numbers. These numbers vary from one problem or from one application to another, but are constants throughout each investigation. A little reflection will show that the presence of such numbers is essential. Something in the law must serve as a measure of those characters that distinguish each individual case. Something must distinguish between one measuring machine and another, between one observer and another, between one rifleman and another.

In Gauss's law there are four of these undetermined numbers. Two of them depend on the position on the target of the center of the pattern; the third and fourth depend on the amount of spreading, or dispersion, vertically and horizontally, and are a measure of the accuracy of rifle and rifleman. Again we come across the standard deviation! As we have already assumed that the center of the pattern coincides with the point aimed at, the first two of these numbers are in fact determined. There remain the third and fourth, which must be found statistically; that is, by actual firing at the target. If this were not so, if these latter numbers were determined by the statistical law, we should know at once that a mistake of some sort had been committed; for merely assuming the existence of accidental errors could not conceivably tell us anything about the marksmanship of the rifleman.

We have not as yet introduced the idea of statistical probability. To do so we can imagine the target to be divided into small squares, or, still simpler, into concentric rings about the center, in accordance with the usual practice in target shooting. In the latter case it is essential to assume that no systematic errors are present. Suppose now that a large number of rounds are fired, say 100,000, and that the number of shots in each of the rings is counted. The statistical probability of hitting any one of the rings is the number of shots in that ring, divided by the total number of shots, 100,000. Suppose that there are six rings (including the bull's-eye) and that the number of shots in the one containing the bull's-eye is 50,000, the numbers for the others being successively 30,000, 10,000, 6,000, 3,000, 1,000. Then the corresponding probabilities are 1/2, 3/10, 1/10, 6/100, 3/100, and 1/100. If the firing is continued, the betting odds are even that a given shot lands in the bull's-eye, 7 to 3 against its landing in the next ring, and similarly for the others. In practice, these probabilities will follow closely those indicated by Gauss's law.

Thus the statistical approach to target shooting has led us to a definite set of chances and so to a prediction of the future fall of the shots. But this prediction will hold good only as long as the gunner holds out; it is subject to all the frailties that his flesh is heir to. He may become fatigued, or get something in his eye, or one or more of a thousand other things may happen to spoil his aim. If we eliminate him altogether and imagine the rifle mechanically held in position, as is done in the munitions plants, our results have to do with the quality of arms and ammunition, instead of arms and

the man, and are less subject to sudden and unpredictable variations.

Let us compare briefly the important features of these two statistical situations that we have used as illustrations: the stature of American males, and the shot pattern on the target. We shall find that they correspond rather closely. To the center of the shot pattern corresponds the average height of American males, as computed from the sample of one hundred thousand individuals. To the scattering of the shots about the center, which we called the dispersion, corresponds the spread of heights about the average. Furthermore, we have seen that the tiny accidental errors responsible for the spreading of the shot pattern are as likely to deflect the shot upward as downward, to the right as to the left, and that as a consequence the grouping of the shots follows Gauss's Law of Errors, a form of statistical law. In the case of the heights of American males, we cannot speak of "accidental errors," for the phrase would have no meaning. But it is nevertheless an *experimental* fact that the grouping of the heights about their average closely follows Gauss's Law of Errors.

When the numbers in a statistical table (arranged as a frequency distribution) follow the Law of Errors, it is called a *normal* or *symmetrical* distribution. The general appearance of such a distribution, illustrated in Figure 1, shows that the values are symmetrically placed about the vertical line that represents the average of all the values, which explains why it is referred to as symmetrical.

Thus the scattering of the shots and the distribution of the heights both follow Gauss's law. But it is to be particularly noticed that while, in the case of the rifle firing, it was possible to say *in advance* of the actual firing that this law would be followed, no such statement could be made in the case of the measurements of heights. In the latter case, this fact was discovered *after* the measurements had been made and entered in the table.

In a simple game of chance, as we have seen, it is possible to predict in advance of any trial or experiment not only what sort of distribution of possible cases will result—that is, what statistical law is followed—but the relative frequency of occurrence, or probability, of each of these cases. This is going one step further than it was possible to go in considering the target shooting. We can therefore arrange the three situations in order of increasing complexity as follows: simple game of chance, target shooting, table of heights.

. . . As examples of the second of these classifications, we have all varieties of gunnery and artillery and the theory of measurements. Under the third classification come the great majority of statistical situations, in insurance, in the sciences, in business.

It would be a great mistake to conclude that, because the distribution of men's heights follows Gauss's law, this is usually the case in similar statistical studies. If we had weighed the hundred thousand individuals, instead of measuring their heights, we should have obtained a distribution that is *not* symmetrical. Such distributions are

called *asymmetrical* or *skew* (see Figure 2).

This fact about weights impresses most people as astonishing. "If heights form a symmetrical table or curve," they say, "why don't weights?" Their feeling is akin to that of the majority of mathematicians and statisticians for several decades following the death of Gauss. Because the exact assumptions required to produce Gauss's law had not yet been completely formulated, and because the authority of the great Gauss hung heavily on them, they tried their best to force all statistical series into the mold of the Gaussian Law of Errors. All true statistical series must be symmetrical, they argued, and concluded that asymmetrical series cannot represent *true* statistical data. They assigned various sources of the difficulty, that the data are inadequately analyzed, or that the number of cases is not large enough, for example. Such a situation does not exist in tossing coins or in throwing dice, where the distribution becomes more and more symmetrical the larger the number of trials. We shall see an example from dice throwing farther on. But isolated instances do not justify the conclusion that all statistical series are symmetrical. The overwhelming majority actually are not.

Let us try to rationalize the situation, at least partially. In obtaining the Law of Errors, which is symmetrical about the average, the assumption was made that the small elementary errors are as likely to be in one direction as in the opposite direction. Suppose that we were to assume, on the contrary, that each error is twice as likely to be in one of the directions, say north, as in the other, south. Then there would be a piling up of errors to the north. The curve would not be symmetrical. This example is sufficient to indicate that the symmetrical is a special case.

All varieties of skew distributions are met with in statistical practice. In a normal distribution (one that follows Gauss's law), more of the values in the table are packed closely about the

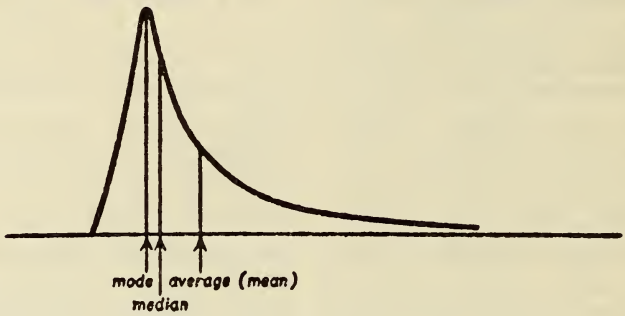


FIGURE 2

SKEW FREQUENCY DISTRIBUTION

The original data represent sales from a mail-order catalog. The curve represents the result of averaging and smoothing the data by the appropriate methods.

average than about any other number. This means that the highest point of the curve shown in Figure 1 corresponds to the average of the table from which the curve was constructed. In a skew distribution the average value and the highest point of the curve, the mode, do not coincide, as indicated in Figure 2.

Perhaps we can best illustrate the important features of a statistical table, or frequency distribution, by using an example from the realm of games of chance, approached this time from the statistical or empirical point of view. Table 1 gives the results of a widely known experiment in dice throwing made by W. F. R. Weldon, which was conducted as follows: Twelve dice were thrown 4,096 times; when the 4, 5, or 6 came up on any die, it was entered as one success, and the number of successes in each throw was recorded. For example, the fourth column of the table reads three successes against 198 throws; this means that on 198 of the 4,096 throws, exactly three of the twelve dice showed a 4, 5, or 6.

TABLE 1

Number of successes:	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	Total
Number of throws:	0	7	60	198	430	731	948	847	536	257	71	11	0	4,096
Theoretical number of throws:	1	12	66	220	495	792	924	792	495	220	66	12	1	4,096

The average number of successes, as computed from the table, is 6.139. We shall pass over, for the time being, the details of this calculation. . . . But we can also compute this average without reference to experiment, since we are dealing with the simple game of dice. It is $1/2 \times 12$, or 6.000; the discrepancy between theory and practice is 0.139.

The proportion of successes, in other words the probability of success, indicated by the experiment is 6.139/12, or 0.512. The theoretical chance of success is evidently 1/2, or 0.500 (three favorable cases out of a total of six cases). The discrepancy between theory and experiment is 0.012.

This comparison between theory and practice, as far as it goes, seems to indicate at least a fairly close accord. But we are far from having a real measure of the agreement between the two sets of figures, and we are not yet prepared to answer the question. Are the results of the Weldon dice-throwing experiment in accord with the laws of chance? The answers to questions of this sort are of the greatest importance in many statistical problems, as we have already seen, and it is time that we indicate in more detail one of the methods that the statistician uses in attacking such problems.

We have studied the Weldon experiment up to this point as a frequency distribution consisting of 4,096 repetitions of a certain event, namely throwing twelve dice at a time. We have

computed the average of the table, which tells us the average number of successes per throw, and from it the statistical probability of success, as indicated by the experiment. Both these quantities, as we have seen, are slightly in excess of the theoretical figures. But we have observed several times, in these pages, that the throw of each die is independent of that of the others, and therefore that throwing twelve dice once is the same thing, as far as the theory of chance is concerned, as throwing one die twelve times. It follows that we can equally well regard the Weldon experiment as consisting of $12 \times 4{,}096$, or 49,152 throws of a single die. As the chance of success on each throw is 1/2, the most probable number of successes is $1/2 \times 49{,}152$, or 24,576. Let us first of all compare this with the actual number of successes observed. To find the latter we must multiply each number in the first line of Table 1 by the corresponding number from the second line. This gives $(0 \times 0) + (1 \times 7) + (2 \times 60) + \ldots$ and so on, and the total is 25,145, which is a deviation from the expected value of 569 in excess.

Our problem is this: Is a deviation of 569, or 2.3 per cent, one that might be expected to occur, according to the laws of chance, reasonably often? . . .

In order to find the answer we must first accurately define and compute the *standard deviation* which, we recall, measures the scattering or dispersion of a set of numbers about its average. As the standard deviation is the square root of a certain average, we define first its square. *The square of the standard deviation is the average of the squares of the deviations of the numbers in the set from their arithmetical average, or mean.* The positive square root of the quantity so computed is the standard deviation itself.

A simple example will make this clear. Suppose that there are four men in a room whose heights are 65, 67, 68, and 72 inches. The average of their heights is 68 inches. The deviations of the heights from this average are respectively -3, -1, 0, and $+4$. The squares of the deviations are $+9$, $+1$, 0, and $+16$, and their sum is 26. Their average is one fourth of 26, or 6.5. The standard deviation is therefore the positive square root of 6.5, which is equal to 2.55.

Now that we know how to compute the value of the standard deviation, we are close to the answer to our question. Our next step requires us to make a new assumption, however, regarding our experimental data. We must assume that the frequency distribution is one that closely follows the normal distribution (Figure 1). If this is not true, the method that we are following requires modification. But if the data represent a series of independent events with constant probability, as in throwing dice or tossing coins, it is proved mathematically that the distribution is close to normal, provided that the number of trials is large, and that the probability of success in each is not too small. We are therefore safe in proceeding with our study of the Weldon experiment.

Under these conditions a quantity known as the *probable error* is of great statistical significance, and it is found from the standard deviation by multiplying the latter by 2/3 (more ac-

curately 0.6745). Its significance and value are due to the following fact: The probable error is a deviation in either direction from the average such that the betting odds that a deviation, selected at random, will exceed it are even. If, then, we find in an experiment like that of Weldon that the total number of successes differs from the expected number by two thirds of the standard deviation, as computed from the ideal frequency table furnished by the theory of the experiments, we shall know that in the long run such a deviation should happen about one time in two. But the odds against a deviation greater than a given multiple of the probable error increase very rapidly with increasing values of the multiple, as shown in the following table.[1]

TABLE 2

Multiple of Probable Error (plus or minus)	*Odds against a Larger Deviation from the Expected Value*
1	1 –1
2	4.5–1
3	21 –1
4	142 –1
5	1,310 –1
6	19,200 –1
7	420,000 –1
8	17,000,000 –1
9	100,000,000 –1

The first line of this table merely repeats the definition of the probable error as that deviation (in either direction) on which it would be fair to bet even money. The second line, however, tells us something new. It states that the probability that a random deviation will exceed (in either direction) *twice* the probable error is 1/5.5 (odds 4.5—1 against). And the last line states that there is only 1 chance in about 100,000,000 that the deviation will exceed *nine* times the probable error.

This table is in reality an extract from a more complete one, that gives the probabilities of deviations of fractional multiples of the probable error as well as whole-number multiples of it. But it is sufficient to show us how very rapidly the probability of a deviation decreases with the size of the deviation.

We need now to know how large the probable error is, as predicted by the theory of chance, for a series of 49,152 throws of a die, the probability of success on each throw being 1/2.

Fortunately, this has been worked out for us by the mathematicians, and we need only take their results and apply them to the case before us. Here is what they have found: If you make a series of any number of trials, say n, each with probability p of success, q of failure $(p + q = 1)$, and if these trials are independent, as they are in rolling dice or tossing coins, the probabilities of various numbers of successes and failures are given by the terms of $(q + p)^n$. To find these terms you multiply $(q + p)$ by itself n times, according to the rules of algebra. The first term is q^n, and it tells you the probability of 0 successes (n failures). The second term is npq^{n-1}, and it tells you the probability of getting exactly 1 success and $n - 1$ failures, and so on for the other terms. Let us illustrate

[1] H. C. Carver, *Handbook of Mathematical Statistics* (H. L. Rietz, Ed.). Houghton Mifflin Company, Boston, 1924, p. 100.

just how this works for a simple example. Suppose that we toss a coin four times, and ask the probabilities of various numbers of heads, so that a success means heads. As we have just learned, these probabilities are given by the terms of $(q + p)^4$. Here both q and p are equal to 1/2, so that we have

$$(1/2 + 1/2)^4 = 1/16 + 1/4 + 3/8 + 1/4 + 1/16.$$

The probabilities we wish are given respectively by the terms on the right side of this equation, as shown in the following table:

No. of Heads	*Probability*
0	1/16
1	1/4
2	3/8
3	1/4
4	1/16

Notice that the sum of the probabilities is 1, as it must be. . . .

Now if you repeat this series of n trials a number of times, say N, the frequencies of occurrence of the various numbers of successes are given by the terms of $N\ (q + p)^n$. This gives us, then, the theoretical frequency distribution that applies to any experiment of the nature of the Weldon experiment, which we are studying. Notice that since $p + q = 1$, the sum of all the terms is exactly N, as it must be.

We wish to know the average and the standard deviation of the frequency distribution given by $N\ (q + p)^n$. Since it is easily shown that the value of N makes no difference, we can as well think of it as 1 and ask the same questions for $(q + p)^n$. The solution of the problem is easily found by mathematical methods, keeping in mind the definitions of the average and the standard deviation, and we shall merely state the results. The average of the distribution, which coincides with what we have called the expected number of successes, is np, and the formula for the standard deviation is $\sqrt{npq}$. Since symbols, not numbers, have been used, these results are general and can be applied to any case, regardless of the values of p, q, and n. We shall appreciate this fact when we apply our findings to the Weldon experiment, for the frequency distribution is then given by the terms of $(1/2 + 1/2)^{49,152}$, and the direct numerical calculation of the average and the standard deviation would therefore involve 49,153 terms. This is a small example of the power of mathematical methods.

We may note, incidentally, that the theoretical distribution given in the third line of Table 1, which applies to the experiment in the form there recorded, is obtained from the terms of $4{,}096\ (1/2 + 1/2)^{12}$. The exponent 12 is the number of dice thrown on each of the 4,096 trials.

Now that we have an expression for the value of the standard deviation, we at once obtain that for the probable error. It is $0.6745\ \sqrt{npq}$.

We are finally in a position to return to the Weldon dice experiment, and to compute the probable error that the theory of chance specifies. According to the above formula it is $0.6745\ \sqrt{49{,}152 \times 1/2 \times 1/2}$, which is equal to 74.8. The chance that a random

deviation from the expected number of successes, 24,576, will be more than 75 is therefore about even. What is the chance that an experiment will result in a deviation of 569, as did that of Weldon? This is a deviation of 7.6 times the probable error, and a glance at Table 2 tells us that the odds against such an occurrence are roughly several million to 1. The actual odds are about 4,500,000 to 1.

What are we to conclude from this extraordinary result? That the laws of chance are incorrect? Certainly not, for in this experiment we have been dealing with mechanical contrivances, namely a set of dice, and in order to be able to test the laws of chance we should have to be assured in advance that the dice were perfectly true, or, at least, they had a perfectly uniform bias, so that the chance of success would be constant. Now the only practical way to test the accuracy of dice, which are always under suspicion due to the varying number of spots on different sides, is to compare the results of rolling them with the results predicted by the laws of chance. We are back precisely to where we started, and we see clearly that experiments in rolling dice can test only the trueness of the dice. We shall have to turn in another direction if we wish to test the laws of chance.

We are still faced with the fact, however, that Table 1, looked at as a frequency distribution, yields an average (and, as a matter of fact, a standard deviation) that is in fairly good accord with the pure chance values. Our conclusion must be that the dice used in the experiment tended to turn up a slightly disproportionate number of fours, fives, and sixes, as a group, but that the *distribution* of successes by trials was in close accord with the laws of chance. A glance at the third line of the table, which gives the theoretical frequency distribution, computed according to the laws of chance, tends to bear out this view. You will notice an excess of throws in which the number of successes is from 6 to 10, inclusive, and a corresponding deficit where the number of successes is from 0 to 5, inclusive. The effect of this is to displace the actual frequency curve, as compared with the theoretical, somewhat to the right (the direction of an increasing number of successes). In other words, the distribution of successes is of the type anticipated, but the probability of success on each trial is not exactly 1/2.[2]

Since mechanical contrivances, coins, dice, roulette wheels, and so on, are subject to mechanical biases, we shall turn, for our next illustration of these

[2] Another of the Weldon dice-throwing experiments, very similar to that studied in the text, is discussed by R. A. Fisher in his *Statistical Methods for Research Workers*, 6th edition, page 67, one of the leading technical works on the subject. In this instance twelve dice were thrown 26,306 times, and on each throw the number of 5's and 6's showing were listed as successes, thus leading to a frequency table altogether similar to our Table 1. The conclusion reached by Mr. Fisher is that the results obtained would happen on the pure chance hypothesis (the dice being assumed true) only once in five million times. This conclusion is so close to that reached in the text that one can suspect that the same set of dice was used in the two experiments. Or possibly all dice have a very small bias, due to the different numbers of spots, which shows up only in very long series of trials.

statistical procedures, to an example of another sort. In drawing cards from a pack the problem of avoiding possible biases is very much simplified, so that we can feel confident that the results obtained by a skillful experimenter should agree with those predicted by the laws of chance. One of the leading modern statisticians, C. V. L. Charlier, made 10,000 drawings from an ordinary deck, recording as a success each drawing of a black card. The total number of black cards drawn was 4,933, giving a deviation of 67 from the expected number, which is of course 5,000. Applying our formula for the probable error we obtain $0.6745 \sqrt{10{,}000 \times 1/2 \times 1/2}$, which equals 0.6745×50, or 33.7. The actual deviation is, then, almost exactly twice the probable error, and Table 2 tells us that we should expect such a deviation about one time in 5.5, in the long run. The result is therefore entirely in accord with that predicted by chance.

This experiment was performed by Charlier with a slightly different object in view. The drawings were recorded in sample sets of ten, count being kept of the number of successes in each.[3] In this way a frequency table was obtained, similar to that in Table 1. Both these tables represent the type of frequency distribution that we have just studied and are known as Bernoulli series. They are characterized by the repetition of independent chance events, the probability of success remaining the same in each. They are so named because the mathematician James Bernoulli made some important discoveries regarding them.

There are many other quantities, in addition to the average and the standard deviation, that can be computed from a table. Each one helps to characterize the table as a whole; each one is shorthand for some important property of the distribution. But it is to be kept in mind that no two, or three, or four characteristics of a statistical table tell its whole story, unless the distribution is one that is known to follow a simple statistical law, a relatively rare occurrence. Even in that event, it is not the table itself that is represented, but the ideal form that the table would assume, if its data were infinitely numerous. This ideal form is in itself of great value, and to find it is one of the objects of many statistical studies. It is of no value, however, until it has been definitely ascertained that the discrepancies between it and the experience table can reasonably be due to chance fluctuations.

The theory of these theoretical frequency distributions is at the root of an important section of mathematical statistics. For the results apply not only to rolling dice, but to any series of trials that are independent in the probability sense of the word, and in which there is a constant probability of success from trial to trial. It is thus possible, by comparing these Bernoullian frequency distributions with others based on empirical statistics, to test the hypothesis that a constant probability lies behind the latter.

In addition to the Bernoullian distribution, or curve, there are two others that are closely related to simple games

[3] The experiment is reported in full in Arne Fisher: *The Mathematical Theory of Probabilities*, 2nd ed., 1930, p. 138.

of chance, and that are of value to the statistician in analyzing statistical material of various sorts. One is known as the Poisson distribution, the other as the Lexis distribution. These three distributions are usually thought of as arising from drawing white and black balls from several urns. We get a Bernoulli distribution if each urn has the same proportion of white and black balls, so that the probability of drawing a white (or black) ball is the same in each urn, and if each ball is replaced before the next drawing from that urn takes place (it comes to the same thing to use a single urn). We get a Poisson distribution if the proportion of white and black balls varies from urn to urn, so that the probability of each varies, and if each ball drawn is replaced at once. We get a Lexis distribution if we make each set of drawings exactly as in the Bernoulli case, but change the proportions of the white and black balls between sets of drawings.

Each of these three cases represents a probability situation that can be solved. . . . On the other hand, the sets of drawings may be thought of as statistical series or frequency tables. In each case the theoretical average and standard deviation can be computed, and they serve as excellent criteria for judging and classifying the statistical series met with in practice. It is clear that in this way the ideas of the theory of chance are brought into the most intimate relation to statistics.

II QUEUING THEORY

✦✦✦✦✦✦✦✦✦✦✦✦✦ QUEUE TIPS FOR MANAGERS

ABE SHUCHMAN

Bottlenecks, backlogs and "jam-ups" are endemic in business. So also is idle capacity. It is, in fact, common to find that in the operation of a production or service facility periods of backlog alternate with periods of partial or complete idleness. On production lines, in the processing of paper work, in receiving materials, in delivering goods, in repairing equipment, and in servicing customers work often piles up. Equipment or personnel cannot meet the demand for service and so units requiring service accumulate. Waiting lines or queues develop, frequently observable at ticket booths, toll gates, bus stops, and supermarket check-out stations. On the other hand, in these very same operations, which are at times

Not previously published.

jammed-up, work to be done often falls short of what can be done. Equipment or personnel available for rendering service become idle and must wait for the arrival of more work. Idle or excess capacity then exists.

Both bottlenecks and idle capacity give rise to costs. These costs originate in the idleness of in-process inventories, equipment and personnel, and in the loss of sales to "impatient" customers. They are not outlay costs but executives are concerned, nevertheless, with their elimination or reduction. Many executives are continually searching, therefore, for that design of a production or service facility which will minimize such costs.

One basic element in the design of a facility is its size or capacity. The problem of design, therefore, frequently confronts executives in the form of questions like: How many furnaces should there be in the heat-treating department? How many clerks should be employed in the billing department? How many check-out stations should the supermarket have? How many saleswomen are needed by the dress department? Questions such as these executives have found, however, are much more easily asked than answered for the knowledge and skill required to find the answers has not been available. As a consequence, executives have had generally to rely on intuition, guess, and hunch or on "cut and try" methods in making decisions about the capacity of a production or service facility.

In many cases such decisions need no longer be made "by guess and by gosh" or by costly experimentation. In recent years mathematicians and statisticians have been studying the process which results in bottlenecks and idle capacity and they have developed a theory which has already been successfully applied in many business and industrial situations. This theory, known as the theory of waiting lines or queues, as the British call them, was fathered by a Danish engineer, A. K. Erlang, who formulated its basic concepts in about 1908 while estimating the amount of central switching equipment needed by his employer, the Copenhagen Telephone Company. His ideas were gradually elaborated during the following four decades, but with very few exceptions, applications were confined to the field of telephony. After World War II, however, workers in Operations Research recognized the applicability of the theory to problems in other fields and the result has been a considerable intensification of interest in the theory and its application to a variety of practical business problems. It is now often possible, therefore, for executives to determine, with accuracy and assurance, the optimum size for a production or service facility.

CHANCE VARIATION

Executives could readily determine the best size for a plant or department if (1) the demand for service was a flow with a regular pattern, and (2) the time required for rendering service was identical for each "customer." Under these conditions all the executive needed to do would be to select the

size which made the rate at which service could be rendered exactly equal to the rate at which "customers" arrived for service. If, for example, "customers" always arrived at fifteen minute intervals and service always required one hour, four service units would obviously be the best number to employ. There would then be no waiting by "customers" and no idleness of service units.

Unfortunately, regularity seldom characterizes production, sales, or service operations. Neither the demand for the product or service of a firm and of its component departments nor the output of a firm and its disparate segments is usually a steady stream with an even flow. Instead, both the rate of demand and the rate of output fluctuate irregularly over time in an unpredictable manner. They are, as the statistician says it, subject to *random* or *chance variation.*

Chance variation in the rate of demand means that there is not always, say, fifteen minutes between the arrival of one "customer" and the arrival of the next. And chance variation in the rate of output means, similarly, that the time used in servicing a "customer" is not always, say, one hour. These times cannot be known beforehand. Sometimes "customers" will follow one another at fifteen minute intervals. At other times, they will follow one another either at longer or shorter intervals. Sometimes, too, "customers" will be serviced in one hour, but at other times servicing will require more than one hour, and at still other times, it will require less than one hour. How much time will be used in servicing any particular customer cannot be known in advance.

NATURE OF THE DESIGN PROBLEM

Such unpredictable irregularities in the demand for service and in the output of a production or service facility make it impossible to match exactly the rate at which service can be rendered and the rate at which "customers" arrive for service. They make it impossible, therefore, to design a facility so that *neither* "jam-ups" nor idleness occur. In other words, whenever the operation of a facility is subject to chance variation, idle capacity or "jam-ups" or both will appear.

Usually, the size of a facility is such that sometimes the number of "customers" who arrive for service during a given time interval exceeds the number that can be serviced immediately. At other times, the number arriving for service falls short of the number that can be serviced without delay. Sometimes, therefore, customers are compelled to wait for service while at other times equipment and personnel are compelled to remain idle while waiting for new arrivals. Thus, chance variation usually results in *both* "jam-ups" and idle capacity.

Of course, the capacity of a facility may be made so large that all "customers" can be serviced immediately no matter what the time spacing of arrivals or the variation in service times. In this case, no "jam-ups" would occur and "customers" would never have to wait for service. None of the costs associated with such waiting

would, therefore, be incurred. However, making the capacity of a facility so large may require not only an enormous investment in equipment or a very large addition to payroll, but will frequently result in a considerable increase in the cost of rendering service. The installation of standby capacity as insurance against the arrival in a cluster of an uncommonly large number of "customers" increases the likelihood that part of the facility will be idle during any given time interval. It results also in an increase in the average proportion of the facility which is idle over any given number of constant time intervals. And the consequence of a rise in the extent to which capacity is underutilized is generally an increase in the unit cost of service.

It is conceivable also that the capacity of a facility may be made so small relative to the demand for its service that idleness of equipment or personnel rarely or never appears. In this case, none of the costs associated with underutilization of capacity would be incurred. However, making capacity so small, increases the likelihood that frequent and severe "jam-ups" will occur. It may very well, therefore, increase the costs arising from the loss of sales to "impatient" customers.

Thus, as capacity is added to a facility the costs due to waiting by "customers" fall, but the costs due to idleness of equipment or personnel rise. Conversely, as the capacity of a facility is reduced, the costs due to idleness fall, but the costs due to "customer" delays rise. In making decisions about the size of a production or service facility, therefore, it is necessary to take account of two kinds of costs that move in opposite directions as the size of the facility is varied.

Now, as we have seen, when the capacity of a facility is small relative to the demand for its service, the costs of idleness in the facility will be low, but the costs of waiting by "customers" will be high. It is likely, however, that the *total cost* of operation of the facility will be high. On the other hand, when the capacity of a facility is large relative to the demand for its service, the costs of waiting by "customers" will be low, but the costs of idleness will be high. Again, the *total cost* of operation is likely to be high. In designing a facility, however, the objective is to find that size which will make the total cost of operation a minimum. The central problem is, therefore, to achieve that coordination of capacity with demand which makes the *sum* of the costs of "customer" waiting and the costs of idleness a minimum.

CHARACTERISTICS OF PRODUCTION AND SERVICE OPERATIONS

A simple production or service operation can be regarded as a system containing three principal elements—an input, a waiting line and servicing units. The input feeds items into the line; the servicing unit takes the items from the line, performs some operation on them and passes them out of the system. For example, consider two workers performing consecutive hand operations on an article, who are sit-

ting side by side at a workbench. The first worker performs his operation and then pushes the article along the bench toward the second worker. The second worker picks up the articles whenever he is ready for one and performs his operation. Here, the first worker is the input, the second worker is the service unit and the articles on the bench between them constitute the waiting line. Given chance variation in the rate at which the first worker pushes articles toward the second worker and chance variation in the rate at which the second worker is ready to pick up another unit, the waiting line between them can contain many or few articles.

Thus, chance variation in arrival and service rates, to use more general terms, causes the waiting line to assume many different lengths. To determine the most economic design for a production or service operation, the proportion of time in which the waiting line is of each possible length must be predicted. To enable such prediction, however, information is needed about the characteristics of the system.

Service operations have four general characteristics relevant to the problem of optimum design about which information is required in order to solve the design problem. These are: arrival times; service times; the number of service units or channels; and the queue discipline.

1. ARRIVAL TIMES

Information is required about the demand for service and this is obtained by observing the times at which "customers" requiring service arrive at the facility. From these observations, one can determine the distribution of the time between arrivals, data which is of particular importance in an analysis of the process resulting in "customer" waiting and idle capacity. Such a distribution is shown in Table 1.

TABLE 1

Time Between "Customer" Arrivals	*Percentage of Arrivals*	*Cumulative Percentage of Arrivals*
0–7.9 hours	86	86
8–15.9	9	95
16–23.9	4	99
24 and over	1	100

It shows, for example, that on the average, 95 of every 100 units arrive after a period of less than 16 hours has elapsed since the preceding unit arrived.

2. SERVICE TIMES

Information is required also about the length of time needed to render service. Sometimes this time will be constant but much more frequently random influences and differing service requirements cause service times to vary. Generally, therefore, information about service times will be in the form of a distribution analogous to that for inter-arrival times. Of course, if the facility contains different kinds of service units, a distribution of service times must be obtained for each kind. For a particular kind of service unit, a distribution of service times might be as follows:

TABLE 2

Service Time	*Percentage of "Customers" Serviced*	*Cumulative Percentage of Units Serviced*
0–1.99 hours	8	8
2–3.99	10	18
4–5.99	60	78
6–7.99	17	95
8 and over	5	100

This table shows, for example, that on the average, 95 of every 100 units are serviced in less than 8 hours.

3. NUMBER OF SERVICE UNITS OR CHANNELS

In analyzing a service operation, the number of service units of each kind in the facility must be known. Also, if units can cooperate in rendering service, as is true, for example, of attendants at filling stations, the amount and type of cooperation that is possible must be known.

4. QUEUE DISCIPLINE

This term refers to the order in which "customers" are selected for service and the way in which "customers" arriving for service behave. Usually, "customers" take their place in a waiting line in the order in which they arrive and are served in the same order. The queue discipline is then "first come, first served." However, there may be other rules governing the selection of the "customer" to be served next. For example, units having certain special characteristics may be given priority or the last "customer" to enter a line may be the first served. Whatever the principle of selection, however, it must be specified. In addition, "customers" arriving for service may not enter a waiting line if it exceeds some specific length when they arrive, or these "customers" may enter the line but lose patience after a time and withdraw. Such behavior has to be measured or estimated. The probability that a "customer" will not enter a waiting line of any given length and the probability that a "customer" will leave a waiting line after any given wait are essential information.

Since each of the characteristics of service operations can assume a number of different forms or values, there are a very great variety of possible queuing situations. For example, arrival times and service times may be constant, normally distributed, or exponentially distributed while the number of service channels may range from one upward. These channels may or may not be able to cooperate. In addition queue discipline may be "first in, first served," "last in, first served," or governed by some priority principle and "customers" arriving for service may be either "patient" or "impatient." There are, consequently, a vast number of different combinations of these characteristics which may be encountered. Many of these combinations have already been analyzed and tables of solutions have been published which can be very helpful when a servicing operation has the same combination of characteristics as one of the models which has already been solved.[1]

[1] See, for example, Peck, L. G. & Hazelwood, R. N., *Finite Queuing Tables*, John Wiley & Sons, Inc., 1958.

A DESIGN PROBLEM

To illustrate the nature and method of solution of a capacity problem, let us consider a highly oversimplified version of a situation which might arise in a firm's customer service department. The situation has the following characteristics:

1. ARRIVAL TIMES

Customer calls for service appear on the average at the rate of 7 per week. The time of arrival of any particular call cannot be predicted exactly as the arrival times are subject to random variation. As a result, any number of calls may arrive in any given week. Over an extended period of time, however, an average of 7 calls may be expected in each work week.

2. SERVICE TIMES

On the average, the time required to reach a customer and render the desired service is five hours. The exact time required to service any particular customer cannot be predicted precisely, however, as the service times vary randomly. Some customers are serviced in very little time, while others require more than the average amount of time. But the time required to service any customer is completely independent of that required by any other customer. Moreover, over the long pull, an average of 8 customers are serviced in a forty hour work week.

3. NUMBER OF SERVICE CHANNELS

The customer service department employs one engineer who serves as a trouble shooter and he is competent to provide any type of service that may be called for.

4. QUEUE DISCIPLINE

To keep the illustration as simple as possible, we assume that each incoming call for service is filed in the order of its appearance and is responded to in the same order. That is, we assume a queue discipline of "first come, first served." Also, once a call comes in, it is never cancelled and since customers do not know how many calls are waiting for service, the length of the waiting line does not discourage them from calling for service.

Suppose now the manager of the department wants to know whether one engineer suffices to meet the demands for service. He is aware that if customers are compelled to wait too long for service, a cost will result, as they will become dissatisfied and take their business elsewhere in the future. He is aware also, however, that in adding engineers so that waiting time is reduced to a minimum, the direct labor costs of his department will increase. His problem is, therefore, to determine the number of engineers that will make the sum of these costs a minimum.

In assuming that both arrival and service times vary completely at random, the illustration has been made an example of a type of waiting line situation which is commonly encountered. Analysis has shown that under these conditions the average length of the waiting line and the average time that a unit requiring service must wait can be readily determined from the average arrival rate and the average service

rate. And knowing this we can proceed without difficulty to the solution of the manager's problem.

Suppose we let A represent the average number of calls per week and S, the average time required to service a customer. Then, the average number of calls waiting for service, N, and the average time a customer must wait for service, W, are given by

$$N = \frac{(A/S)^2}{1 - \frac{A}{S}} \quad \text{or} \quad \frac{A^2}{S(S - A)} \qquad (1)$$

and

$$W = \frac{A}{S(S - A)}. \qquad (2)$$

Accordingly, with the average rate of arrival as 7 calls per week or 1.4 calls a day and the average rate of service as 8 calls per week or 1.6 calls per day, the average number of customers waiting for service is

$$N = \frac{1.96}{1.6(1.6 - 1.4)} = \frac{1.96}{.32}$$
$$= 6.13 \text{ customers,}$$

and the average time that a customer must wait for service is

$$W = \frac{1.4}{1.6(1.6 - 1.4)} = \frac{1.4}{.32}$$
$$= 4.38 \text{ working days.}$$

Similar computations can be made for two engineers, three engineers, and so on. These computations involve the use of formulas which take into account the existence of multiple service channels. Such formulas are readily available, however, and although they are much more complex than those given above, their use poses no problem for a competent statistician. In addition, there are other formulas with which it is possible to compute the probability that any given customer will have to wait for service one day or more, two days or more, etc., with any given number of channels (engineers) available for rendering service.

Thus the manager of our customer service department can be provided with information which tells him how long a customer will have to wait for service, on the average, when one, two, or three, or any number of engineers are available. He can also be told the probability that any particular customer will have to wait for service any given length of time or longer, with any given number of service channels available. Obviously, then, if the department manager has in mind some standard of service which he desires to provide, he can readily determine how many engineers must be employed to provide it. Conversely, if his budget does not permit him to employ any more engineers, he can tell his superiors what level of service they may expect his department to render with the personnel already employed.

THE COMMON SENSE SOLUTION

Most executives faced with the problem of designing a service facility are likely to assume that to obtain most efficient operation they must make the average service rate approximate as closely as possible the average arrival rate. This would seem to be the dictate of common sense. In fact, however, a design that makes the ratio of these

rates, referred to as the "traffic intensity," equal or very nearly equal to 1 will ordinarily be far from most efficient when arrivals and service times are subject to chance variation.

To see this, let us examine the extreme case in which A/S, the "traffic intensity," is exactly equal to 1. If this value is inserted in formula (1) above, the formula for determining the average length of a waiting line when both arrival and service times are random, we get the "astonishing" result that the average number of units waiting for service is infinite. Moreover, it can be shown further, as will be done shortly, that the smaller the difference between the average arrival rate and the average service rate, the greater will be the average number of "customers" who have to wait for service and the longer each will have to wait, on the average. This means that as the average arrival rate approaches the average service rate, the costs of idleness in a facility will fall, but the costs of waiting will rise, and in fact, will rise very rapidly. If these latter costs are significant, then the operation of the facility will tend to become extremely inefficient. The implication here is that when arrival and service times are random, greater economy of operation can usually be obtained by keeping the ratio between the arrival and service rates less than 1, even though this means that service units must be idle some of the time.

As a demonstration of these conclusions, consider the following table which gives the average length of the waiting line and average waiting time for a number of different arrival and service rates, when there is only one service channel and both arrival and service times are random.

TABLE 3

Average Arrival Rate (A)	1	4	6	9	14	15	31	32
Average Service Rate (S)	4	8	9	12	16	16	32	32
A/S	.25	.50	.67	.75	.88	.94	.97	1.00
S/A	4.00	2.00	1.50	1.33	1.14	1.07	1.03	1.00
Average Length of Waiting Line (N)	.083	.50	1.30	3.00	6.13	14.06	30.03	infinite
Average Waiting Time (W) (as a percentage of the time interval used in measuring A and S)	.08	.13	.22	.33	.44	.94	.97	infinite

It can be seen from the table that when the average service rate becomes less than $1\frac{1}{3}$ times the average arrival rate, both the average number of "customers" in the queue and the average time a "customer" must wait for service increase rapidly. At the extreme,

when the average arrival and average service rates are equal, the average length of the waiting line and the average waiting time become infinite.

These observed results arise from the random variation in arrival and service times. For, when the average arrival rate and average service rate are very close and arrival times are random, a cluster of units may arrive at intervals less than the average service time. It is unlikely that the less than average interval between arrivals will be completely compensated for by identical variations in service times, even though these times too vary randomly. As a consequence, a waiting line will be formed. This waiting line is difficult to "work off." At some time, in fact, another cluster of arrivals will appear before a waiting line has been eliminated and these new arrivals will result in a longer line than had been produced by the previous cluster. "Working off" the longer line will be even more difficult and before it is accomplished, still another cluster may appear. As this goes on, short waiting lines arise less frequently and long waiting lines, more frequently. The average of the waiting line and the average waiting time tend to become infinite.

METHODS OF SOLUTION

Solving a capacity problem is almost never as easy as the illustration above makes it seem. The basic data in capacity problems are the distributions of arrival times and service times and these may assume a very great variety of forms. Sometimes these distributions are in a form which permits the development of equations that describe accurately the cumulative distribution of time intervals between arrivals and the cumulative distribution of service times. When such equations can be written, they can be used to deduce formulas similar to those used above but usually much more complex. These formulas can then be used to solve for average length of waiting line, average waiting time, the probability that there are a given number of units waiting in line when the next unit arrives, the probability that a given unit will have to wait some given length of time, and so forth. When such equations can be written, therefore, the capacity problem can be solved by analytical or mathematical methods.

Often, however, no equations can be formulated which describe with sufficient accuracy the relationship between the percentage of arrivals and the time between arrivals or that between the percentage of units serviced and service times. At times, also, although such equations can be written, they are so complex that their use in solving the problem is inordinately difficult and time consuming. But neither of these circumstances need prevent the solution of a capacity problem. The analytical or mathematical methods for solving the problem cannot be used but recourse may be made to another method—the Monte Carlo Method.

THE MONTE CARLO METHOD

The Monte Carlo Method is a technique for simulating the behavior of a service facility. Given the basic data of the process that results in "customer"

waiting and idle capacity—the distributions of arrival and service times, the number of service channels, and the queue discipline—an artificial history of the facility's operation is constructed by means of a random sampling procedure. This history provides a picture of what may be expected to happen during actual operations of the facility. This picture will be only an approximation of the results of actual operation, but if the sampling procedure is repeated, say, several thousand times, the approximation will usually be very close to the actual.

Thus, the average waiting time, the average length of the waiting line, the average idle time of service channels, and so forth can be determined even when no formulas are available for describing the relationships underlying the operation of a facility. In addition, the Monte Carlo Method has another advantage. It permits experimentation. It permits manipulation of those characteristics of the waiting line process which are subject to control and the determination of the effects of any change. For example, the number of service channels can be increased or decreased, more or less time can be allowed for service, and the arrival rate or queue discipline can be altered. And the consequences of any one or any combination of such changes will become apparent in the artificial history of the facility's operation. These consequences can, in fact, be determined very quickly by putting the facts of the situation on a digital computer and "running off," in a few minutes, thousands of arrivals and servicings.

As an illustration of the nature of the Monte Carlo Method and its use in designing service facilities, consider the following highly oversimplified example.

A small plant employing fifty mechanics has a tool crib attended by two men. Sometimes these crib attendants are idle and sometimes more than two mechanics appear at the crib at such short intervals that some must wait for service. The plant superintendent wants to know whether two attendants are sufficient for the most economical operation of the crib and asks a member of his staff to study the operation.

The analyst begins by making sample observations of the crib's operation for periods of, say, one-half hour, at various times over a week or two. He tabulates and analyzes these observations and obtains the following information.

1. There is one chance in five that a mechanic will arrive at the crib for service within any given five minute interval.
2. The time between arrivals is, on the average, three and one-half minutes.
3. The average service time is five minutes, but 40% of the calls at the crib require 4 minutes for service, 30% require 5 minutes, 20% require 6 minutes, and 10% require 7 minutes.

Using this information, the analyst can now simulate the crib operation and from the results determine whether two is the best number of attendants to have.

To estimate how many calls for service will appear in any given five minute interval, the analyst turns to a table of random numbers and selects 50 numbers. These 50 numbers, each of which may, of course, have any value between

0 and 9 represent the 50 mechanics in the plant. Before selecting the numbers, the analyst has arbitrarily decided to let numbers 2 and 3 represent the arrival of a mechanic. Since there is only one chance in 10 of getting a 2 for each number looked at, and similarly, only one chance in 10 of getting a 3 for each number looked at, the chances are 1 in 5 of getting either a 2 or 3 for each number looked at.

The number of 2's and 3's that appear among the 50 selected numbers is assumed to represent the number of arrivals in the first 5 minute interval. If, for example, three 2's and one 3 appear among these 50 numbers, 4 mechanics have come to the tool crib.

The analyst now turns his attention to service times. Once again he refers to the table of random numbers and selects 4 numbers, one to represent the service time for each of the mechanics now assumed to be at the crib. On the basis of the information he has regarding the distribution of service times, the analyst has already decided that numbers 1, 2, 3, and 4 will represent a service time of four minutes; numbers 5, 6, and 7 will represent a service time of five minutes; numbers 8 and 9 will represent a service time of six minutes; and number 0 will represent a service time of seven minutes. In other words, using his knowledge that service requiring four minutes occurs twice as often as service requiring six minutes and four times as often as service requiring seven minutes, the analyst has distributed the numbers between 0 and 9 in the same proportions.

Suppose now that the four numbers just selected from the table of random numbers were a three, a four, a nine, and a seven, in that order. These are assumed to mean then that the first mechanic to arrive at the tool crib required four minutes for service, the second mechanic also required four minutes, the third required six minutes, and the fourth required five minutes.

The analyst has now simulated the operation of the tool crib for one 5 minute interval. He knows how many mechanics have arrived for service and how many minutes it will take to service them. He now repeats the process of drawing random numbers and continues to do so until many five minute intervals have been simulated.

As the analyst proceeds, he finds that waiting lines develop from time to time as more mechanics arrive than the crib attendants can serve immediately. He carefully notes the time that each mechanic must wait in line before being served and when he has run through the desired number of 5 minute intervals, he totals this waiting time. This total is then multiplied by the hourly rate of the mechanics to obtain the cost of time "lost" while waiting at the crib for service, when the crib has only two attendants.

The analyst now simulates the crib operation as it would be with three attendants and then as it would be with four attendants. He obtains, thus, the cost of time "lost" when operating with three and four as well as with two attendants. And given the hourly rate of the crib attendants, the analyst can proceed to determine the sum of the cost of "lost" time and the cost of attendants when either two, three, or four attendants are employed. The in-

formation which he can then give the plant superintendent might look like this:

Number of attendants	2	3	4
Average arrival rate (5 minute interval)	1.43	1.43	1.43
Average service rate (5 minute interval)	1	1	1
Average time "lost" in waiting	5 minutes & 12 seconds	40.5 seconds	7.5 seconds
Average number of arrivals per 8 hr. shift	137	137	137
Average time "lost" per 8 hr. shift (in hours)	11.9	1.54	.29
Mechanics hourly rate	$5.00	$5.00	$5.00
Crib attendant's hourly rate	$2.00	$2.00	$2.00
Average cost of time "lost" per 8 hr. shift	$59.50	$7.70	$1.45
Cost of crib attendants per 8 hr. shift	$32.00	$48.00	$64.00
Average total cost per 8 hr. shift	$91.50	$55.70	$65.45

From this information, the plant superintendent can readily see that the minimum total cost—including both the cost of "lost" mechanics' time and the cost of attendants—is achieved when three attendants are employed. The superintendent can also estimate easily the annual savings which will result from operating with the most efficient number of attendants. These costs and annual savings, it must be emphasized, are "expected" costs and savings. They are the averages that should appear over a considerable period of the crib's operation, with two, three, or four attendants and with the given wage rates.

In this highly simplified example many factors which are often important in waiting line problems have been purposely ignored. Such factors might include variations in arrival and service rates by season of the year, day of the week, time of the day or weather conditions and variations produced by cooperation among servicing units. They might also include absenteeism and overtime. All such factors can be taken into account, however, and the most efficient design for a facility defined by using the Monte Carlo Method. Also, although the discussion of our example noted that the analyst drew ran-

dom numbers and interpreted their meaning, totaled waiting times, and so forth, efficient and economic use of the Monte Carlo Method is usually secured by employing an electronic computer for these tasks. This is especially true when the problem involves complicating conditions such as those just mentioned, for no analyst can, of course, perform these operations as rapidly as an appropriately programmed electronic computer.

SOME APPLICATIONS OF QUEUING THEORY

Waiting line or queuing theory has already been applied to a great variety of situations in business and industry and a brief description of some of these applications may be useful in suggesting problems which the theory may help solve.

1. The New York Port Authority has used queuing theory to analyze delays at the toll booths of bridges and tunnels. The result was a recommendation concerning the number and scheduling of toll collectors and the number of toll booths required at any time of day to provide a given level of service at minimum cost.

2. At the Boeing Airplane Company, foremen complained that their men were waiting too long in lines at tool crib counters. Plant executives considered assigning more attendants but they were under pressure to reduce overhead. Queuing analysis was used, as a consequence, to solve the problem.

3. Several firms have used the theory in attacking the problem of machine breakdown and repair. In such a problem, the machines which break down form, in effect, a line waiting for repairs by the men who service them. It is desired to employ that number of repairmen which makes the sum of the cost of the production loss from "down" time and the payroll cost of repairmen a minimum.

4. A large chain of supermarkets has used queuing to determine the number of check-out stations needed to secure smooth and economic operation of each store at various times during the day and week.

5. Queuing theory has been applied to the design of terminal facilities for ships and trucks. The problem which has been successfully solved is that of determining the number of docks to be constructed. Since both dock costs and the costs of demurrage can be very great, with the latter cost decreasing as the former mounts and vice versa, it is desired to construct that number of docks which will minimize the sum of these costs.

6. Arthur D. Little, a prominent management consulting firm, has used queuing theory in studying a wage incentive plan. In the plant studied, some workers had been assigned to operate two machines while others had been assigned to operate four machines. All machines were identical so all operators were paid the same base rate but the incentive bonus for production in excess of quota was half as much per unit for operators with four machines as for operators of two machines. The study revealed, however, that while each of the two machines run by one man would be "down" about 10% of its scheduled operating time, each of

four machines run by a man in that group would be "down" about 14% of its scheduled operating time. The reason for this was that in the four machine batteries, two or more machines might break down at once and so one or more would have to wait for the others to be repaired. As a consequence, the operator of a four-unit battery had to operate at higher efficiency than the operator of a two-unit battery in order to earn the same incentive wage. The problem was solved by paying operators of four machines at a higher base rate determined by using the probabilities computed from the queuing theory.

Other applications of queuing theory include the timing of traffic signals, restaurant service, car wash service, the scheduling of patients in clinics, the number of switch engines to use in a railroad classification yard, the staffing of a clerical operation, the balance of material flow in a job shop and the design of inventory and production control systems.

CONCLUSION

It appears, then, that queuing theory can be an important aid to executives faced with the problem of determining the optimum design for a production or service facility. It can enable executives to cope, systematically and rationally, with the uncertainties characteristic of such problems. Whenever recorded experience can be made to yield accurate information about the pattern or distribution of arrival and service times, analytical or Monte Carlo Methods can be used to convert this information into the probabilities associated with waiting and idleness of any duration. If realistic values can then be assigned to the costs of waiting and the costs of idleness, the size of staff or physical facilities required to provide service either of a specified standard or at a minimum cost can be defined. The theory of waiting lines or queues is, therefore, a powerful addition to the arsenal of managerial techniques.

III DECISION THEORY

✦✦✦✦✦✦✦✦✦✦✦✦ STATISTICAL AIDS TO DECISION MAKING

CHARLES A. BICKING

INTRODUCTION

This article decribes a direct application of probability concepts to management decisions. The basic principles used are similar to those applied in statistical quality control and design of experiment at

Industrial Quality Control, *August 1958, 7–12.*

operating and technical levels of business. A general pattern of analyzing problems, described popularly as *Design for Decision,*[1] has been applied to real examples of decision making.

The aim is to combine careful estimates of costs and of returns with equally carefully estimated probabilities of the occurrence of various outcomes of alternative courses of action. When costs, returns and probabilities are combined, the value or desirability of the alternatives are expressed in quantitative terms. A comparison of the desirabilities of the several possible courses of action enables the manager to choose the most favorable one.

AN EXAMPLE OF DESIGN FOR DECISION

For example, design for decision may be applied to the consideration of a proffered Government contract. The alternative courses of action are to accept or to reject the contract. The outcomes of either acceptance or rejection may be success for the company in achieving the objectives of the contract, success for a competitor, or failure for the company. In order to come to the most economic decision, it is necessary to assign a probability of occurrence figure to each outcome and to determine, in dollars, the cost and return expected of each outcome. When these figures have been combined to give dollar values, a decision may be reached on the basis of maximizing the company's gain.

Reference to Table 1 will make it possible to describe the detailed steps by which the decision was reached to accept a contract for the development of metallic-fibre reinforced ceramics. The construction of the table follows,

TABLE 1

DECISION DESIGN—METALLIC-FIBRE REINFORCED CERAMICS PROJECT

Alternatives	*Accept*			*Reject*		
Outcomes	*Company Success*	*Competitor's Success*	*Company Failure*	*Company Success*	*Competitor's Success*	*Company Failure*
Probability of Outcome, p	0.6	0.5	0.4	0.4	0.6	0.6
Return (in thousands of dollars)	$P + Y_1 + Z = 130$	$P - Y_2 = -20$	$P = 5$	$Y_1 - X = -50$	$-Y_2 - X - Z = -200$	$-X = -100$
Value or Desirability (p × Return)	+78	−10	+2	−20	−120	−60

Where P = Profit on Contract = $5,000
Y_1 = Value of Commercial Market = $50,000
Y_2 = Value of Loss of Competitive Position = $25,000
Z = Value of Government Production Contract = $75,000
X = Cost of Research = $100,000

row by row, the steps in analyzing the problem.

The first step is to decide what alternatives are open to the company. After deciding that the only alternatives open are acceptance or rejection of the proposal, the conceivable outcomes have to be listed. The relationships of the outcomes determine the way that probabilities are assigned. A successful outcome means establishment of a market for metallic-fibre reinforced ceramics. This market is presumed to be zero right now and would continue to be zero until research has shown that the material can be produced commercially to compete with other materials that it might replace.

In this instance, company success and company failure are conditional. Either one or the other outcome must occur and the sum of the probabilities must be 1.0. It is considered, however, that the probability of a competitor's success is independent of the company's success; that is, the sum of all three probabilities under a given alternative will be greater than 1.0. It is possible to conceive of a situation in which the probabilities of all three of these outcomes would be conditional. For instance if it is likely that the second comer will abandon the field to the first comer. Another distinction to be made here is whether success means simply coming up with a satisfactory product or whether it means following through to a commanding position in the market. It should be pointed out in connection with the outcomes of rejection that the company planned to do the work anyway, even without contract support.

The assignment of probability values requires considerable experience and background. Although it is not usually necessary to estimate these precisely, it is quite important that the relative magnitude of the assigned probabilities be correct. The more information the administrator has about the technical aspects of the project, about the capacities and limitations of his company and its competitors, and about the field of work in general, the better the assignment of probabilities will be. What this means is that the decision can be no better than the information and skill that goes into making it. The design merely insures that all the available skill and information are used. It permits more clear-cut separation of many possible outcomes. It must be remembered the design is for the use of the administrator and does not relieve him to delegate decision to inferiors in knowledge and judgment. The design is a tool by which a skillful administrator marshals and surveys his facts.

In the example, the probability of company success upon acceptance of the contract is modestly set at 0.6 and it follows that the probability of company failure must be 0.4. The competitor, who is a smart fellow, too, but will not have the advantage of government sponsorship, is assigned a slightly smaller probability of success, 0.5. Upon rejection, the company's outlook would be more pessimistic, due to inability to assign as much effort to the project, and a low probability of 0.4 is assigned. Consequently, the probability of company failure is 0.6 under this alternative. The competitor, however, with government support, is assumed to have the same advantage the com-

pany would have had, and is assigned a probability of 0.6 of success.

It is now necessary to arrive at some dollar figures for costs and returns. The profit from the project will be small. The value of the commercial market, considered for some relatively short payoff period, say three years, is about ten times as large as the profit. The loss of competitive position if a competitor also succeeds in developing the product is worth about half the total commercial value. The value of having the inside track on government production contracts is 50 percent greater than that of the commercial market (also on a short time scale). However, the cost of the research is large, of the order of 20 times the profit.

If the project is accepted and the company succeeds, the return will be the profit ($5,000), plus the value of the commercial market ($50,000), plus the value of government production contracts ($75,000), or $130,000. If a competitor succeeds, the return to the company will be the profit ($5,000), minus the value of the loss in competitive position (—$25,000), or —$20,000. If the company fails it will have only the profit from the contract ($5,000).

For the alternative of rejection, if the company succeeds, it will have the commercial market ($50,000), minus the cost of the research (—$100,000), or —$50,000. If a competitor succeeds, the return will be zero, minus the value of the lost market (—$25,000), minus the cost of research (—$100,000), minus the value of the government contract (—$75,000), or —$200,000. Finally, if the company fails, the return will be zero, minus the cost of the research, or —$100,000.

When these returns are multiplied by the appropriate probabilities, the value or desirability of each outcome is known. The most desirable outcome is the one with the largest plus value. Accordingly, analysis of this design has indicated that the company should accept the proposal.

The above example has shown the steps followed in arriving at a decision, but much still remains to be said about the underlying concepts and the decision rules that may be followed. Before going further into these matters, however, the steps in the decision-making process will be summarized. This has been done in Table 2.

TABLE 2

STEPS IN DECISION MAKING

1. List the possible alternative decisions.
2. List the possible outcomes of these decisions.
3. Choose a decision rule, focusing attention on the short or the long run, on loss-control or maximizing expected gain, depending on the business climate.
4. Rank or assign probabilities to the outcomes.
5. Estimate costs and returns and either rank or assign dollar values to each.
6. Evaluate each outcome using the chosen decision rule.
7. Test the value or desirability against some policy standard.

The decision to accept the proffered Government contract was reached by applying a simple decision rule: the outcome was selected which had the

largest plus value. This is but one of a number of rules suggested in the literature for application under differing circumstances.

A DECISION REGARDING CAPITAL EXPENDITURE

The operation of some of these rules may be illustrated by applying them to a decision to approve or disapprove a request for a capital expenditure for purchase and installation of a stone saw for cutting ball wheels. This equipment was to be used for cutting sections out of ball wheels to meet customer specifications. The wheel finishing department had no available machinery to perform this operation. Part of the work was being done in the Abrasive Engineering Laboratory and part in the Machine Shop. It was claimed that the installation would reduce manufacturing cost, relieve the Abrasive Engineering Laboratory and the Machine Shop of production operations, and substantially improve delivery schedules through reduced transportation and handling time.

The rules will be applied successively to this problem and Table 3 will be built up a step at a time.

The alternative outcomes of approval of the request are: obtaining the advantages claimed, obtaining some advantages (but less than claimed) and failing to obtain any advantages. Careful consideration of the request and its accompanying letter of justification led to arranging the outcomes thus, in descending order of probability of occurrence. This arrangement is not inconsistent with a rather conservative appraisal of the request, since anything greater than a 50-50 chance for the first outcome should result in the above order. The outcomes of disapproval, in descending order, are: failing to obtain the advantages claimed, incurring further losses, and losing the business altogether. The possibility of further losses and even, eventually, of losing the business are inherent in the situation. However, they are remote as long as the business is vigorously handled. By far the most probable outcome of disapproval is maintaining the status quo. For the time being, we need go no further in constructing Table 3a.

APPLICATION OF VARIOUS DECISION RULES

One of the simplest rules calls for choosing the alternative for which the return of the most probable outcome is largest. This does not require numerical evaluation of probabilities nor of returns. It requires only the ranking of the returns under the most probable outcomes for each alternative. It emphasizes the probabilities and is often used when one outcome has a very high probability. If the request is approved, the most probable outcome is that the advantages claimed will be obtained: if disapproved, that the advantages will not be obtained. Obviously, there is no need to refer to the dollar returns in the table because the former outcome will have the larger return except in the unlikely situation that the long term return is no larger than the cost. Therefore, the decision would be made to approve the request. Any desired policy limitation may be applied to qualify

this decision, such as the very reasonable one that the return per annum shall not be less than 25 percent of the cost. On this basis, this decision is which takes into account interest factors. The effect of interest rates is implicit if not explicit in the use of average annual returns in this example.

TABLE 3

DECISION DESIGN—PURCHASE OF STONE CUTTING SAW

	Alternatives	*Approve*			*Disapprove*		
a. Decision Based on Probability Only	*Outcomes*	*Obtain advantages claimed*	*Obtain less than claimed advantage*	*Fail to obtain advantages*	*Fail to obtain advantages*	*Incur Further Losses*	*Lose the Business*
	Probability of Outcome—In Order of Size	(1)	(2)	(3)	(1)	(2)	(3)
	Cost—$	−6,570	−6,570	−6,570	0	0	0
	Return *—$	22,350	11,175	0	0	−12,375	−67,500
	Net	15,780	4,605	−6,570	0	−12,375	−67,500
b. Decision Based on Desirability	Probability of Outcome, p	0.5	0.4	0.1	0.6	0.3	0.1
	Cost—$	−6,570	−6,570	−6,570	0	0	0
	Return—$	22,350	11,175	0	0	−12,375	−67,500
	Net	15,780	4,605	−6,570	0	−12,375	−67,500
	Value or Desirability (p × Net Return)—$	7,890	1,842	−657	0	−3,713	−6,750
c. Decision Based on Variable Probability	Outcome, p Probability of	0.4—0.6	0.3—0.5	0—0.15	0.5—0.7	0.2—0.4	0.1—0.2

* Over useful life of equipment, 15 years.

borderline, since the return claimed per annum (1/15 of $22,350) is only 22.7 per cent of the cost ($6,570).

It is common practice to make decisions on the basis of present value Computation of savings has taken into account cost of money for new capital expenditures. In a more general way, the application of a policy limitation to the number of years allowed for a

pay-off reflects the degree of concern for variations in the economic climate.

Further discussion of the dollar returns used in the example will follow later.

A second rule calls for choosing the alternative which *could* lead to the largest return. This is a very optimistic rule. It focuses on the returns. It does not require evaluation of probabilities. Here we could again make our selection without referring to the numerical size of the returns although it is helpful to have them because of the policy qualification applying. It is enough to know which is the largest. The decision is the same as under the first rule because the largest net return ($15,780) is expected for the outcome of obtaining the advantages claimed upon approval of the request.

Under a third rule we would consider the least favorable return under each alternative and choose the alternative most favorable of the two. This is a very pessimistic rule. It emphasizes security. It applies what is known as a "loss-control criterion." It does not require evaluation of probabilities. This rule calls for comparing the return if approval fails to result in obtaining any advantages with the return if disapproval results in loss of the business. It does not require dollar figures to conclude that under this rule, the request should be approved. Note that this minimizes the loss (–$6,570 compared to –$67,500) if the worst should occur under either alternative. (Availability of the dollar return figures permits realistic appraisal of the worst eventualities.) This rule avoids application of limiting policy and is a panic solution.

To apply a fourth rule, it is necessary to complete Table 3. This rule uses both probabilities and returns. It is applicable whether the probabilities are independent or contingent. It is the rule applied to the earlier example.

In Table 3b contingent probability values have been assigned to the outcome under each alternative. Trust in the validity of the estimates of cost and return is only moderate, so that an even chance of obtaining the advantages claimed is expected. There is almost as high a probability that some advantage will be obtained though not as much as claimed, so a probability of 0.4 is assigned. The probability of failing to obtain any advantages if the proposal is approved is 0.1, by difference. If the proposal is disapproved, the probability of obtaining none of the claimed advantages is quite high, so 0.6 is assigned. Similarly, there is a pretty good chance, 0.3, that some further loss will be incurred and a small though real probability, 0.1 (obtained by difference), that the business will be lost altogether.

The cost is based on a proposal from a manufacturer of stone and marble cutting machinery and a Works Engineering estimate of installation labor and materials. The Industrial Engineering Department estimated that installation of this unit would reduce direct operating costs by 50 per cent. On the basis of this and a 15 year depreciation period, the annual savings were calculated to be $1,490. Over the 15 year period, therefore, the return will

be 15 × $1,490, or $22,350. The actual return could, of course, be larger or it could be zero. It is likely that the actual return might be about half the amount claimed or $11,175, and this amount is assigned to the second outcome. The return on the third outcome is zero.

If the proposal is rejected, there will be no return. It is quite likely that due to poor service further losses would be incurred, say to the extent of one-tenth of the orders and 100 per cent increase in service costs. Assuming an average value of $150.00 per wheel, a demand of 300 wheels per annum, and operating income of 15 per cent of sales, the value of the business is $6,750 per annum. Ten per cent of this (—$675), plus the additional service cost (—$150), times 15 equals —$12,375, which is the loss entered in Table 3b. It is assumed that the business might be lost, not at once, but after five years, so that the loss given in the table for this outcome is —$6,750 times (15–5), or —$67,500.

In applying the fourth rule, the net returns are multiplied by the probabilities to give the last line in Table 3b, the value or desirability. The largest value is $7,890, which leads us to the same decision as all the preceding rules. We would again, as a final step, apply the policy requirement concerning ratio of the first year's return to cost.

Rule number five requires the calculation of the mathematical expectations of the two alternatives. This rule uses both probabilities and returns. It applies when probabilities are contingent. It focuses on what happens in the long run. For the alternative of approval, this is:

$$\begin{aligned} E_A &= (0.5 \times \$22{,}350) \\ &\quad + (0.4 \times \$11{,}175) \\ &\quad + (0 - \$6{,}570) \\ &= \$11{,}175 + \$4{,}470 - \$6{,}570 \\ &= \$\ 9{,}075 \end{aligned}$$

For the alternative of disapproval, this is:

$$\begin{aligned} E_D &= 0 + (0.3 \times -\$12{,}375) \\ &\quad + (0.1 \times -\$67{,}500) \\ &= -\$\ 3{,}713 - \$6{,}750 \\ &= -\$10{,}463 \end{aligned}$$

Note that, where the probabilities are conditional, this amounts to adding up the values under the respective alternatives. The higher mathematical expectation points to the choice of the alternative of approval. Here again, some policy restriction regarding significant differences in expectancies might reasonably be applied. It is suggested that an appropriate limitation might be that the difference between the highest and lowest expectations, reduced to an annual basis, shall not be less than 25 per cent of the cost. For the example,

$$[\$9{,}075 - (-\$10{,}463)]/15 = \$1{,}303$$

which is 20 per cent of the cost ($6,570). Therefore, the decision to approve is marginal according to this rule.

A sixth and final rule makes it possible to arrive at a decision combining loss-control with mathematical expectation when reliable information about the probabilities is not available. This rule can be applied in the absence of reliable information about probabilities

(i.e., a range of probabilities may be used). It combines loss-control with mathematical expectation. Table 3c is modified to show a range of probabilities under each outcome.

The mathematical expectations may be calculated using the lower bound of the most probable outcome, the upper bound of the least probable outcome and adjustment of other contingent probabilities, within their bounds, to give a total probability of 1.0. If approved, the mathematical expectation of the value is:

$$\begin{aligned} E_A &= (0.4 \times \$22{,}350) \\ &\quad + (0.15 \times 0) \\ &\quad + (0.45 \times \$11{,}175) \\ &\quad - \$6{,}570 \\ &= \$8{,}940 + \$5{,}029 - \$6{,}570 \\ &= \$7{,}399 \end{aligned}$$

If disapproved:

$$\begin{aligned} E_D &= (0.5 \times 0) \\ &\quad + (0.2 \times -\$67{,}500) \\ &\quad + (0.3 \times -\$12{,}375) \\ &= -\$13{,}500 - \$3{,}713 \\ &= -\$17{,}213 \end{aligned}$$

The higher expectancy for approval indicates that it is the alternative that should be accepted. Applying the usual policy qualification, as before,

$$[\$7{,}399 - (-\$17{,}213)]/15 = \$1{,}641,$$

which is just 25 per cent of the cost. The proposal would be approved as the basis of providing a satisfactory return on investment and protecting the company against possible losses.

This final rule appears to be a very satisfactory one for dealing with decisions of the sort considered. It employs probability concepts without requiring extremely fine determination of the probability of each. It introduces control of losses, which is not taken into account when only the positive alternative outcomes of a decision are considered. It implies most complete utilization of the available data having a bearing on the decision.

Having a rule like this and applying it conscientiously has other secondary advantages. By focusing attention on the estimation of probabilities, costs, and returns, application of the rule should increase the ability to make valid estimates of these elements of decision. In other words, the very existence of rules and a system for applying them should aid in improving the data fed into the design. It emphasizes the necessity for doing things about getting progressively better and better data.

The six rules that have been illustrated are tabulated in Table 4 to summarize this section and to make them readily available for reference.

PROOF OF THE DATA BY QUALITY CONTROL

If it turned out that the cost and savings estimates used in the decision design procedure did not come from controlled data sources, the method of solution itself would be suspect. The proof of the data by statistical quality control is, therefore, one of the most important steps that needs to be taken.

Fortunately, since the subject of administrative applications of quality control has received wide attention, methods for determining whether the data are controlled have become well

TABLE 4

SOME FREQUENTLY APPLIED DECISION RULES *

1. Choose the alternative for which the return of the most probable outcome is largest.
2. Choose the alternative which *could* lead to the largest return.
3. Choose the alternative with the most favorable of the least favorable returns under the respective alternatives.
4. Choose the alternative for which an outcome has the largest numerical value or desirability.
5. Choose the alternative with the largest mathematical expectation.
6. Choose the alternative associated with the largest of the least favorable expectations.

* With the exception of Rule Number 4, these are from the work Bross: *Design for Decision*, The Macmillan Company, New York, 1953.

known. Follow-up procedures for improving uncontrolled estimates will depend upon the nature of the records kept and of management practices in use. This is a large but manageable problem, which could provide material for not one, but several articles. It is assumed that before the decision design procedures are applied, steps will be taken to assure that the necessary conditions of control in the data have been met.

One specific way in which statistics can help to improve the basic data used in decision making is through the application of the concept of statistical control charts to the difference between estimates of costs and returns and the actual costs and returns obtained. The availability of historical data for this purpose depends on the existence of good records, particularly of unit costs. The manager, faced with making a decision, needs to have confidence that the estimating procedures are in control in a statistical sense. He will, therefore, have to insist on maintenance of adequate records and on the employment of statistical quality control principles to eliminate the occurrence of erratic estimates. This is not difficult to do, once its value is understood. For example, electrical construction estimates have been studied by control chart means.[2] The differences between estimates and actual costs for a large number of jobs were tabulated. When these data were plotted on control charts, it was possible to determine the degree to which the work of the estimators was inconsistent, or not controlled. Also, the limits could be used to determine how well acceptable performance standards were being met. In that study, the differences between estimates made by the central engineering department and by plant engineers were clearly pointed out. Also, reduced variability between estimate and performance as the result of introduction of a manual of standard unit cost data could be measured.

Approximately one hundred separate estimates were available, divided equally among material estimates and labor estimates. There was also approximately equal division among plant engineering department estimates and central engineering department estimates, both before and after introduction of the standard cost manual. Only

the plant estimates showed lack of control, and then only on the range chart. This indicated lack of consistency between different estimators.

However, the most important information given by the charts was that although inconsistent practices resulted in very wide limits in all instances, there was a very marked decrease in the spread of limits from plant estimates to central engineering estimates, and from central engineering estimates before to estimates from the same source after publication of the standard

WORK SHEET—DECISION DESIGN

Project:	*Subject:*					
Alternatives						
Outcomes						
Probability, p						
Return, $						
Cost, $						
Value, $						
Decision						

1. Useful life:
2. Return during useful life:
3. Detail of Probability and Return:
4.
5.
6.
7.
8.
9.
10.
11.
12. Expectations:
13. Alternative:
14.
15. Alternative:
16.
17. Sum of Expectations ÷ Useful Life:
18. Line 17 + Cost:
19. Remarks:

costs. In terms of the upper limit for the average charts, with plant estimates indexed as 100, the results were as follows:

Plant Estimates: 100
Central Engineering Estimates
Without Unit Costs: 72
With Unit Costs: 30

CONCLUSION

In making decisions of the type described, managers have generally supplemented meager data with the use of intuitive judgment. Intuition may be fortified by the scientific handling of quantitative data. This fortification may take the form of systematic "designs for decision." The use of a systematic approach focuses attention on the estimation of probabilities of success and failure and on the validity of estimates of costs and savings. Improvement in the estimation of probabilities is the real crux of this matter. It is achieved through unrelenting effort to approach reality, as well as through experience.

In general, great confidence is put in the validity of engineering estimates of costs. Sometimes these can get out of line, but with reasonable care and adherence to sound managerial policies a satisfactory result is usually achieved. The validity of estimates of savings is somewhat more open to question if only because of a very natural tendency to make a proposal appear as attractive as possible. It is suggested that quality control charts may be used to determine the consistency of differences between estimates and actual savings. It is worthwhile to consider means of making information on actual savings more available. A routine follow-up using quality control methods will make it possible to find the assignable causes of exceptional deviations between estimates and experience and lead to progressive improvement in the correctness of estimates and hence of decisions based on them.

The techniques recommended may appear to be complex. However, the effort required to use them is small compared to the economic value to the company and the enhanced reputation of the decision-maker that would ensue were statistical aids utilized fully.

REFERENCES

(1) Irwin D. J. Bross, *Design for Decision,* The Macmillan Company, New York, 1953

(2) Charles A. Bicking, "The Control Chart as a Tool in Management," paper presented at 104th Annual Meeting of the American Statistical Association, Washington, D.C., December 29, 1944

ACKNOWLEDGMENT

Acknowledgment is made of the generous assistance given by Maj. Gen. (U.S. Army, Ret.) L. E. Simon, Director of Research and Development, and R. S. Bingham, Jr., Supervising Quality Control Engineer, both of The Carborundum Company, in developing these applications.

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE LOGIC OF QUANTITATIVE DECISIONS

DAVID W. MILLER

One of the most interesting developments of the past fifteen years in the area of analytical tools for executives has been that of a field known as decision theory. Decision theory undertakes to consider the structure of quantitative decisions in themselves, abstracted from any specific problem which may have occasioned the decision problem and abstracted from any specific involvement in particular quantities. In short, decision theory is concerned with what the medieval scholastics would have called the quiddity of the quantitative decision process. As such, the conclusions of decision theory underlie most of the powerful and well known methods of quantitative analysis of business decision problems—statistics and operations research are cases in point. In this article we propose to discuss some of the ideas and conclusions of decision theory with regard to the rational decision making process.

The basic structure of decisions is quickly discoverable—and has been used by many disciplines other than decision theory. Philosophers, sociologists, psychologists, economists, management theorists, and many others have had occasion to utilize this basic structure so we can briefly recapitulate the major features of the typical decision structure. First is the fact that we do not usually refer to a "decision" unless the decision maker has, in fact, various alternative courses of action. Even the law will recognize that a man with a pistol to his head may not have alternative courses of action and, hence, cannot be considered to have made a decision. We shall refer to the various available alternative courses of action in any decision problem as the strategies of the decision maker. We assume that the decision maker can select any one of his available strategies at his option. Since the selection of a strategy is under the control of the decision maker it is sometimes convenient to use mathematical terminology and say that the strategies are based on the variables which are controllable.

Second, it is an equally obvious fact that reality is rarely so tractable as to make the result or outcome of a selection of strategy depend solely on the strategy selected. In other words, the outside world will usually play a crucial role in determining what happens as a result of the selection of strategy. In some way, the outcome which results will depend both on the selection of strategy and on what happens in the outside world. Using mathematical terminology again we are saying simply that there are usually some uncon-

Not previously published.

trollable variables which also have an effect on the result of any specific selection of strategy. It is as if the world or "Nature" could also select a strategy in some game against the decision maker—and we shall return to this concept subsequently. In any event, we need some descriptive phrase to describe the world's side of the decision problem. We shall refer to the various possible things which can occur in the world and which are relevant to the outcome of the selection of a strategy as states of nature.

So far, then, our picture of the decision problem is that the decision maker has various strategies, any one of which he can select, and that there are various alternative states of nature which may occur in the world. The selection of any specific strategy and the occurrence of any specific state of nature will result in some outcome. A particularly convenient way to portray this structure is in terms of an outcome matrix: the rows are the various strategies, the columns are the various possible states of nature, and the outcomes are given at the intersection of the corresponding columns and rows. An outcome matrix, then, would look like this:

	States of Nature				
Strategies	N1	N2	N3	N4	...
S1	O11	O12	O13	O14	...
S2	O21	O22	O23	O24	...
S3	O31	O32	O33	O34	...
...	.	.	.	.	...

The O's, of course, represent the outcomes.

Let us try this skeleton out on the flesh of a specific decision problem. Consider an inventory decision. The strategies are the various amounts which may be ordered of the item in question. The states of nature will include, at least, the various possible levels of demand. The outcomes, finally, will consist of the results for each combination of a given amount ordered and a given level of demand. Thus, for example, if twenty units are ordered and if demand is for 14 units the outcome will be that 14 units are sold and six units are left in stock. At this very basic level our simple structure seems to work reasonably well.

A third, and last, key component of the decision structure remains to be introduced. This is that the decision maker has objectives, ends, goals, purposes, or any other synonym of these words, which he wishes to achieve and which are his motivation for making the decision in the first place. Such objectives can be as various as the men who make decisions but there is always at least one objective for any decision problem. The objective may be to make the maximum dollars profit, or to accomplish some purpose with the minimum dollars cost, or to achieve the maximum degree of peace and quiet, or to achieve the largest possible market share, or to retain control in a proxy fight, or to retire at the smallest possible age, or any one of an unlimited number of other possible objectives. But whatever the specific objective or objectives involved in a given decision problem, the central notion from our point of view is that each outcome in the outcome matrix has, in terms of the objective or objectives, a worth. The

worth of the outcome is the degree to which the outcome contributes to the achievement of the objective in question. Frequently we can express the worth of each outcome in some numerical form. When this can be done we say that the number representing the worth of a specific outcome is the payoff for that strategy and that state of nature. This granted, we can, then, equally well present the payoff matrix for a decision problem:

	States of Nature				
Strategies	N1	N2	N3	N4	...
S1	P11	P12	P13	P14	...
S2	P21	P22	P23	P24	...
S3	P31	P32	P33	P34	...
...	.	.	.	.	...

The payoff matrix is the basis of the decision theory analysis of quantitative decision problems. The adjective, "quantitative," refers precisely to the fact that the worth of the outcome is measurable. By definition, the numerical measure of this worth is called the payoff.

Let us return to our inventory decision problem and consider what form the payoffs might take. Clearly this depends on the specific objective involved in the inventory decision problem. However, the commonest objective for inventory problems is the achievement of the maximum possible profit. In this event it would generally be easy to convert the outcomes to payoffs. We would be satisfied to determine the dollar worth of each outcome and to use this dollar worth as the payoff for the given strategy and state of nature. Using our earlier numerical example, if we sell fourteen units we would make the profit per unit times fourteen. The six units left in stock might, for example, have to be scrapped and would have a known per unit scrap value. These two components could be combined to give the net profit (or loss) which resulted from the given outcome. This, then, would be the payoff for the strategy of ordering twenty units and the state of nature representing a demand for fourteen units.

Obviously, a most important question is: When can the worth of the outcomes be numerically measured? In other words, when can the outcome matrix be converted to a payoff matrix? Complicated arguments from the theory of measurement are involved in any detailed answer to this question. However, since our purpose here is primarily expository we will try to avoid these complexities by confining our attention to some major kinds of decision problems for which this problem is not an important one. It is certainly the case that for a very great number of ordinary business decision problems the objective is either the achievement of profit or, what amounts to another way of looking at the same thing, the avoidance of costs. For such decision problems as these dollars is ordinarily a satisfactory measure of payoff. For this kind of decision problem the measurement problem does not ordinarily arise so we will confine our attention to this case. However, the importance of the measurement problem in general justifies two paragraphs devoted to the other kinds of decision problems.

The greatest difficulty in converting

outcomes to payoffs arises for decision problems for which the objective is simply not one for which there are natural numerical measures. For example, if an executive wants to achieve a relative degree of peace and quiet it is difficult to see how the worth of outcomes can be measured. Similarly, how can the worth of the outcomes in a proxy fight be measured? For such cases as these we often cannot achieve numerical measures in the usual sense. However, it is sometimes possible to rank the outcomes in order of worth and this is sufficient for some kinds of analysis. In other cases it is possible to use a technique called the standard gamble in order to achieve a numerical measure of the worth of outcomes. The standard gamble was created for this purpose by Von Neumann and Morgenstern (Theory of Games and Economic Behavior; Princeton University Press, 1947) and the interested reader can find a discussion of the technique in Luce and Raiffa (Games and Decisions, Wiley, 1958). Our point here is to emphasize that there are some methods available for the measurement of payoffs in cases other than the one we will be considering.

A different kind of difficulty can arise even in decision problems for which dollars would appear to be a natural measure of the payoffs. This problem is related to the basic economic concept of utility. It may be the case that the decision maker's utility for a given number of dollars is not measured by the number of dollars. When this is the case it follows that the payoff cannot be measured by amounts of dollars since these would not reflect the true worth of the outcomes to the decision maker. That this may be the case is evident from such an everyday occurrence as taking insurance. Obviously, the utility of the insurance company for the various possible outcomes must be different from that of the person taking insurance else both parties would not be content with the transaction. Nonetheless, both the insurer and the insured reason in terms of dollar amounts. As a general rule, we can say that if the various outcomes involve amounts of money which are marginal with respect to the total resources of the decision maker, then it will usually be the case that the amount of money will measure the utility of the amount of money. If some of the outcomes involve amounts of money which represent appreciable percentages of the total resources of the decision maker, then it will probably be the case that the amount of money will not represent the utility to the decision maker of that amount. In this latter case it is possible to use the standard gamble technique to achieve numerical measures of the actual utility of the various outcomes. Therefore, we do have means for handling this kind of case. However, we will not further consider this possibility here.

Granted, then, the payoff matrix representation of a specific decision problem, what does decision theory have to say about the decision process? The first major statement of decision theory is that it is necessary to classify decision problems in accordance with what the decision maker knows about the likelihood of occurrence of the various possible states of nature. There are

four major kinds of decision problems with regard to this classification. The first class is called decision problems under certainty. For these problems there is, in effect, only one possible state of nature. More specifically, these are problems for which the outcome, and the payoff, is determined solely by the decision maker's selection of a strategy. The outside world plays no part in the determination of the outcome. In other words, the payoff matrix has, effectively, only one column. At first consideration such a decision problem would hardly seem to represent a real problem at all. Suppose the objective is to minimize costs. Then for each strategy there would be one and only one outcome and we would know the dollar cost of the strategy. Therefore, the decision maker would only have to select that strategy which had the smallest cost of all of his available strategies.

With regard to this reasonable argument we may note two things. First, we have surreptitiously introduced a second major concept from decision theory. After a specific decision problem has been represented in payoff matrix form it is necessary for the decision maker to select a strategy. This, after all, is the whole point of the decision process. In decision theory this is translated to mean that some decision criterion must be applied to the payoff matrix which will lead the decision maker to select that strategy which is most in accord with his goal of achieving his objectives. A decision criterion, then, is some procedure for selecting one strategy from the available possibilities after the decision problem has been represented in payoff matrix form. Part of our above argument concerning decision problems under certainty, then, is that the decision criterion for this kind of problem is obvious: select that strategy for which the single possible payoff is most in accord with the objectives—the largest payoff for profits or the smallest payoff for costs.

Second, our argument is absolutely right, as far as it goes. There doesn't seem to be any problem for this kind of decision "problem." However, the argument overlooks one possibility. This is that for some kinds of decision problems the number of strategies may be so enormous that it is absolutely impossible to evaluate the payoffs for each one of them. For example, consider a company which has five plants and which produces 1,000 different items. Let us assume that any plant can produce any item. Generally, there will be different unit costs involved in producing any given item in any particular plant. For a production schedule for a month, say, the company will certainly want to produce all of its items as cheaply as possible. Granted the correctness of the company's cost accounting data, this is a decision problem under certainty since there is one specific, and calculable, payoff for each strategy. A strategy is any particular way of assigning all the products to the plants so that the desired amounts are produced. Unfortunately, however, there are an inconceivably huge number of strategies for this problem: 7 followed by about 3,500 zeros. If all of the largest computers in existence were put to work on this problem simultaneously it would take just this side of

forever to evaluate all of the payoffs for this decision problem. Therefore, there can be a very serious problem of resolving a decision problem under certainty. Fortunately, there are a variety of mathematical procedures which permit us to solve such gargantuan decision problems without having to actually evaluate all of the payoffs. These are the operations research methods called programming: linear programming, dynamic programming, quadratic programming, and so forth. For our present purposes, however, it is sufficient to note that there is no problem about the decision criterion for this kind of decision problem.

The second major class of decision problems is called decision problems under conflict. The defining characteristic of this kind of decision problem is that the decision maker is faced with one or more rational opponents—competitors in the business world, for example. On first consideration this class of decision problems seems to be in ill accord with our basis for classification: the knowledge of the decision maker concerning the likelihood of occurrence of the various possible states of nature. (The states of nature for this kind of decision problem are, of course, the strategies available to our opponent or opponents.) However, we do have, in this case, a specific kind of knowledge about the likelihood of occurrence of each of the competitive strategies: we know that our rational opponent will select his strategy with the intention of achieving as much as he can for himself. It turns out that this insight is a very powerful one. This class of decision problems is the subject of the theory of games—originated by Von Neumann and Morgenstern in the book previously referred to. The rich and fascinating subject of game theory is beyond the compass of this article but, once again, it certainly merits a few paragraphs.

The major problem of game theory is the same problem which we will be facing subsequently: the determination of a decision criterion. The basic question in game theory is: What is the degree of conflict of interest between the opponents? It turns out that this depends, at least in part, on two important dichotomies by means of which we can talk about four kinds of games. The simplest case is where there is complete conflict of interest and there are only two opponents. The classic example of this kind of game is any two-person parlor gambling game, such as gin rummy. This, as a matter of fact, accounts for the title of the theory, "game theory," even though most of the "games" of life are matters of deadly urgency like wars, diplomacy, and nuclear negotiations with Russia. In a game like gin rummy there is complete conflict of interest between the opponents because anything which is to be won by one player must be lost by the other player. Such games are called zero-sum games since the sum of the payoffs of any such game is always zero: if A wins \$5 then B must have lost \$5 and $\$5 - \$5 = 0$. Gin rummy, then, is a two-person zerosum game.

Other parlor games are zerosum but have more than two players. Poker is a

classic example of this kind of game, an n-person zerosum game. While there is complete conflict of interest somewhere in such a game it is not as simple as a two-person game. This is because coalitions of players can and do form in n-person games. In other words, one group of players may form an explicit or tacit coalition against one or more other players. Such a coalition often forms quite automatically in a poker game, for example, when the losers form some kind of coalition, united by their misery, and vent their spite on the winners, if they get the chance. The defining characteristic of a coalition is that the players somehow jointly select their strategies to maximize the returns to the coalition as a whole. As such, there is no longer complete conflict of interest among the players in a coalition. In game theory further distinctions are made such as whether there is or is not the possibility of side payments among the members of a coalition. However, we must rest content with the observation that lack of complete conflict of interest enormously complicates the analysis of a game. It does this precisely by making it difficult to discover a valid decision criterion. Therefore, n-person zerosum game theory is much less definitively developed than is two-person zerosum game theory.

The same problems can arise in two-person games which are not zerosum. This is typical of most games outside the parlor. As a matter of fact, it is doubtful that there is any game-type situation in the real world which has complete conflict of interest. Consider, for example, the negotiations between a company and the union representing its work force. This is a two-person game kind of situation but it is hardly zerosum, no matter what it may appear on the surface. The reason for this is that the two parties in this game have a mutual desire to maintain company operations. Naturally, each side wants to maintain operations on its terms, and in this fact lies the conflict of interest. However, each side would benefit by maintaining operations as compared to terminating them, and in this fact lies the lack of complete conflict of interest. Most actual two-person games are similar in this respect—there is not complete conflict of interest. It turns out that this fact causes grievous difficulties with regard to finding a satisfactory decision criterion. Obviously, the case of n-person nonzerosum games is worse since both difficulties are compounded.

For the case of the two-person zerosum game, however, there is a complete theory and it is generally agreed that there is only one decision criterion which a rational decision maker can use in this kind of decision problem. It will be worth considering the argument which leads to this criterion. Suppose we have the following payoff matrix for a two-person zerosum game:

		Player B		
		T1	T2	T3
Player A	S1	−4	−2	5
	S2	3	−1	2
	S3	6	−4	−2

Here the T's designate B's possible strategies. There is no necessity, of course, that the two players have the same number of possible strategies. The payoffs are in dollars. Precisely because the game is zerosum it is only necessary to present one payoff matrix since what one player wins, the other loses. Here the payoffs are presented in terms of A. Thus, if A selects his S1 and B selects his T2 the payoff is —$2. This means that A loses $2 and, obviously, that B gains $2. In order to analyze this decision problem we will introduce one virtually self-evident idea; the idea of one strategy dominating another strategy. Compare, for example, B's strategies T2 and T3. Remembering that B wants to make the payoffs as small as possible, it is clear that B will always do better with his T2 than with his T3, no matter what A does. Thus, if A chooses his S1 then B will win $2 with his T2 whereas he would lose $5 with his T3. If A chooses his S2 then B will win $1 with his T2 whereas he would lose $2 with his T3. And, finally, if A chooses his S3 then B will win $4 with his T2 whereas he would win only $2 with his T3. Thus, B does better with T2 than he does with T3 for any strategy which A might select. We say, then, that T2 dominates T3. Furthermore, the virtually self-evident idea is that it will never be to a rational player's advantage to use a dominated strategy, T3 in the present case. He would always do worse than he need do if he uses a dominated strategy and, therefore, he will never use one. Thus, B will not use his T3 and the game payoff matrix is really only

		Player B	
		T1	T2
Player A	S1	−4	−2
	S2	3	−1
	S3	6	−4

But now we note that A's S2 dominates his S1, since A wants to make the payoffs as large as possible. Therefore, A will never use his S1 and the payoff matrix is really

		Player B	
		T1	T2
Player A	S2	3	−1
	S3	6	−4

We can repeat the reasoning again. This time B's T2 evidently dominates his T1 and then, in the remaining matrix, A's S2 dominates his S3. Therefore, the payoff matrix finally boils down to

	Player B *T2*
Player A S2	−1

The original payoff matrix has been reduced to one in which there is only one strategy for each player. The optimal strategy for each is, therefore, this one strategy. The worth of the game to A is —$1. This means that he will lose $1 every time that the game is played. However, this is the best he can possibly do in this game. This analysis has been accomplished through the one idea of dominance and the

ity distribution governing the occurrence of the states of nature is known. This means simply that the probability of occurrence of each state of nature is known. Thus, consider a simple inventory problem. Let us suppose that demand for a given item can never exceed five units. A probability distribution of demand for this item might look like this:

Demand	*Probability*
0	.10
1	.25
2	.35
3	.15
4	.10
5	.05
	1.00

The sum of the probabilities of the various states of nature must add up to one because some one of the states of nature must occur. The demand probability distribution shows the probability of occurrence of each possible level of demand. Thus, the probability of occurrence of a demand for three units is .15. This means that if the same situation were repeated a great many times there would be a demand for three units about 15% of the time. The inventory decision problem for this item would be one under risk because we know the probability distribution governing the occurrence of the states of nature. Whenever we know this probability distribution, and only when we know it, we are dealing with decision making under risk.

Now, for decision problems under risk there is a general consensus that there is a specific decision criterion which rational persons should use. Before we proceed to present it let us give an example which we will use throughout our subsequent discussion. Instead of taking a business decision problem which would require some more or less involved discussion in order to present, evaluate the payoffs, and defend the presentation let us take a very simple kind of situation which virtually everyone has either experienced or at least knows about. This example has to do with the situation facing anyone who is at a racetrack and who is considering making a wager on a specific horse in a specific race. Such a person has four possible strategies: don't bet, bet that the horse will win, bet that the horse will place (run second), and bet that the horse will show (run third). Similarly, there are four relevant states of nature: the horse will win, place, show, or lose. The minimum bet is $2 and the winnings, if there are any, depend on the total amount bet on that horse and on that race in a relatively complicated way. However, let us assume that we know what the various winning amounts would be. Then the payoff matrix might be:

		States of Nature			
		Win	*Place*	*Show*	*Lose*
STRATEGIES	Don't bet	0	0	0	0
	Bet win	8	−2	−2	−2
	Bet place	2	5	−2	−2
	Bet show	1	2	3	−2

The various payoffs are in rough accord with the usual experience in such cases. Now, the question in this, as in any other, decision problem is: What strategy should be selected? In decision theory terms this requires the utilization of some decision criterion. Under the assumption that this is a case of decision making under risk it follows that the probabilities of the various possible states of nature must be known. We will reserve comment on the question of how, or whether, these probabilities might be known in the present case. Instead, we will simply assume that the probabilities are known to be:

State of Nature	*Probability*
Win	.30
Place	.50
Show	.15
Lose	.05

The knowledge of these probabilities makes the decision problem one under risk.

Now, the generally accepted criterion for decision making under risk, mentioned earlier, is the use of expected values, or averages. According to this criterion the average payoff should be calculated for each strategy and that strategy should be selected which has the largest such expected, or average, payoff. Before we argue for and against this criterion let us illustrate how it would be used on our example. The calculation of the expected payoffs for the various strategies is very easy. All that is required is the multiplication of each payoff for the strategy by the probability of the corresponding state of nature and the addition of the products so obtained. This is the expected payoff for the strategy. Thus, for the strategy of not betting we have simply $.3(0) + .5(0) + .15(0) + .05(0) = 0$. The expected value of the payoff for this strategy is 0, as would be expected. For the other three strategies we calculate:

Bet win	:	$.3(8) + .5(-2) + .15(-2) + .05(-2) = 1.0$
Bet place	:	$.3(2) + .5(5) + .15(-2) + .05(-2) = 2.3$
Bet show	:	$.3(1) + .5(2) + .15(3) + .05(-2) = 1.65.$

Thus, the strategy of betting place, with an expected payoff of $2.30, has the largest expected payoff. In accord with the expected value criterion this is the strategy which should be selected by the decision maker.

What is the argument for this criterion? Assume that this identical decision problem were repeated a very great many times. Then it can be shown that if each time that strategy with the largest expected value were selected that the total returns over all of the decisions would be larger than they would be from any other selection of strategies. In particular, for our example, on the average the decision maker would net $2.30 for each such decision problem for which he bet place. An immediate argument against the criterion arises. No decision problem is ever exactly repeated. Why, then, base your selection of a strategy on a series of hypothetical events which will

never happen—namely, the repetition of this decision problem? This argument can be answered. It can be shown that for any series of decision problems under risk the total return will be the largest if the decision maker selects each time that strategy which has the largest expected value. A new counter-argument now is presented. Suppose the decision maker doesn't have sufficient resources to undertake a series of such decisions since bad luck early in the series will wipe him out. What reason is there, then, to base the selection of a strategy on a hypothetical series of events which will never happen? Fortunately, we have already answered this objection. We earlier made the proviso that dollars could not be used to measure payoffs if the dollar amounts represented an appreciable proportion of the decision maker's total resources. So if this last counter-argument is valid the payoffs were incorrectly stated and the utility of the decision maker for the dollar amounts should have been measured instead. It will be remembered that we suggested that one device for accomplishing this was the standard gamble technique. Now, it can be shown that if the standard gamble technique is correctly used for this purpose, then the resulting payoffs will be such that the use of the expected value as a decision criterion will be in accord with the decision maker's objectives and, hence, his optimal procedure. Therefore, we conclude that the arguments against the use of the expected value criterion can be answered and that this is the criterion which should be used in decision making under risk.

The fourth, and last, class of decision problems is, perhaps, the most interesting from our present point of view. This class is called decision making under uncertainty. This class of decision problems includes all those cases where the probabilities of occurrence of the various states of nature are not known. In particular, it would generally include our particular example of racetrack betting. Many important business decision problems fall into this class but before we can indicate what kinds of decision problems these are it is necessary to take a short side trip.

In many fields of thought there are some deep underlying disputes which verge on being metaphysical in nature. Mathematics is a case in point. In the foundations of mathematics there is a dispute raging between intuitionists and formalists as to just what is and what is not permissible in mathematical arguments. Nonetheless, this long-continuing argument has no effect on any practical applications of mathematics which we may have occasion to make. This is so because both sides in the dispute are in agreement by the time they get to the kind of mathematics which we would be likely to have occasion to apply. Therefore, at the level of applications there is really no dispute. We are not so fortunate in probability theory. Here there is an equivalently basic dispute—and one which is very much older than the one in mathematics. It still rages with the same bitterness after several hundred years of argument. Unfortunately, however, the dispute in probability theory has to do with the very meaning of probabilities and it is very close to the surface. Therefore, we cannot avoid it when we

want to apply probability theory. In particular, it is of considerable importance in our present context. The argument has two opposing sides: the objectivists and the subjectivists. While we cannot hope to present the details of this quite fascinating dispute it will be worth while to indicate the basic question at issue and how it is relevant to our present concerns.

Oversimplifying somewhat ruthlessly —but remaining true to the issue—the objectivists maintain that we can only talk about probabilities when we can construct frequency tables. In other words, we can talk about the probability of a coin showing heads on a toss because we can toss the coin many times and make a frequency distribution of the number of times it shows heads and tails. We may be able to talk about the probability distribution of demand, for example, because we can use statistical and econometric procedures to determine the frequency distribution of past demand. But we certainly couldn't talk about the probability of our horse winning the race because this identical race will only be run once and there is no possibility of determining a frequency distribution with regard to our horse winning the race. A business decision problem might have as relevant states of nature such things as peace, war, depression, recession, and so forth. The objectivist would deny the possibility of meaningfully talking about probability distributions in this context because it would be impossible to determine a frequency distribution. Speaking very roughly, we can say that most statisticians are objectivists. Objectivism has been the dominant school in probability theory for several decades.

The subjectivists agree with the objectivists that if a frequency distribution is available it will generally supply the probabilities of the relevant states of nature. However, they maintain that probabilities can be used in a different way. This is as measures of the degree of belief of the decision maker in the likelihood of occurrence of the various states of nature. They propose to measure such degrees of belief in such a way that the resulting measures will behave exactly like probabilities. Furthermore, they have developed methods for actually measuring degrees of belief in this sense. In other words, the subjectivists maintain that most decision problems which the objectivist would consider to be under uncertainty can be converted to ones under risk by measuring the degrees of belief of the decision maker. Now, as will be seen subsequently, there is no generally accepted decision criterion for decision problems under uncertainty. This severely limits the analytical possibilities for this kind of decision making. By the objectivist standards, an enormous number of practical business decision problems would fall into the uncertainty category. This means that for these business decision problems there would be very little help which the analyst could offer the executive. For the subjectivist, on the contrary, most of these decision problems could be converted into ones under risk and a number of analytical possibilities would then be available for the assistance of the decision maker. Now, this in itself is no argument in favor of sub-

jectivism because the objectivist can simply answer that if a problem can't be solved, it can't be solved and that is the end of the matter. However, there are two answers to this argument. First, the origin of the dispute is almost only a philosophical predilection and this seems an insufficient basis for denying the possibility of the utilization of powerful analytical tools. Second, the objectivist is really maintaining that if there is no frequency distribution then there is no usable information about the probability distribution governing the states of nature. This would seem to deny the obvious fact that any competent decision maker has a wealth of information concerning the states of nature relevant to his decision problems. The subjectivist proposes simply to make this information explicit and to put it in a usable form. Partly for these reasons there has been a resurgence of subjectivism in the past decade. Books such as Leonard Savage's "Foundations of Statistics" and Robert Schlaifer's "Probability and Statistics In Business Decisions" have presented powerful statements of the subjectivist position. The writer is himself a subjectivist but for the remainder of this article we will take the objectivist position. This is necessary since for the subjectivist there is no particular problem connected with decision making under uncertainty.

The problem of decision making under uncertainty is simply one of finding an adequate decision criterion. The trouble is that there is an embarrassment of riches. There are a number of decision criteria, each having some arguments in its favor. We may begin our discussion by considering one proposed by Wald. Wald suggested that we should act as if we were faced with a rational opponent in a two-person zero-sum game. In other words, we should use the Wald criterion discussed previously. This is also known as the pessimistic criterion or the criterion of complete conservatism since the decision maker pessimistically, or conservatively, assumes that the worst will happen to him, no matter what strategy he selects. The application of this criterion is, of course, straightforward. For our decision problem the worst that can happen if no bet is made is a return of 0. For any other strategy we can lose $2. Therefore, the maximin strategy is not to bet and this is the strategy which the Wald criterion would dictate. We note for future reference that this is the way this criterion is usually applied but that it is not really quite correct. Before we assume that a pure strategy is called for we should investigate and see whether the payoff matrix has an equilibrium point or not. In this case it does but we shall return to a consideration of this point below.

Now, of course, when we are feeling moody we are quite likely to assume that "Nature" will deliberately make it rain if we plan a picnic or make the snow melt if we plan to go skiing. However, in our more rational moments we know that this rather egocentric assumption is something less than a valid one. But Wald was not really suggesting such an anthropocentric approach to the decision problem. His argument is that if we really don't know anything about the probability of occurrence of the various states of nature we must

protect ourselves against the worst. Opponents of this criterion point out such payoff matrixes as this one:

	N1	*N2*	*N3*	*N4*	*N5*	*N6*
S1	−1	100	100	100	100	100
S2	0	0	0	0	0	0

For this decision problem the Wald criterion would dictate the selection of S2 and, so the opponents say, this is clearly illogical. Supporters of the criterion reply that if nothing is known about the probabilities then it is possible that the probability of N1 is .999-99999999 and that in this case it wouldn't be illogical at all. So the debate goes on, with no conclusion likely to be reached. However, we shall offer a criticism of this criterion below which will at least indicate that it is not reasonable in the important class of decision problems we have been concentrating on in this article.

Some opponents of the Wald criterion have complained that there is no reason to concentrate attention completely on the worst that can happen for each strategy. Suppose, for example, that the decision maker is a complete optimist instead of a complete pessimist. Reasoning analogically he should choose the maximax strategy. Since the best he could do would be to win $8 if he bet win and the horse won, it would follow that this should be the strategy selected by the complete optimist. No one has defended such a decision criterion but Hurwicz has suggested a criterion which permits the decision maker to be rational if he feels lucky. This criterion involves the use of a quantity which Hurwicz named the coefficient of optimism. The interested reader can find a procedure for measuring the coefficient of optimism in Luce and Raiffa's "Decisions and Games." Here we will be satisfied to describe its use. The coefficient of optimism is a number between zero and one and it is the weight that will be ascribed to the best possible payoff for each strategy in determining the optimal strategy. One minus the coefficient of optimism will be the weight ascribed to the worst possible payoff. Hurwicz's criterion demands the calculation of the weighted average of the best and of the worst payoff for each strategy. The strategy with the largest such weighted average is the strategy selected by the Hurwicz criterion. Suppose the decision maker in our example has a coefficient of optimism of 1/2. To apply the Hurwicz criterion we would calculate:

Strategy	*Best*	*Worst*	*Average*
Don't bet	0	0	$\frac{1}{2}(0) + \frac{1}{2}(0) = 0$
Bet win	8	−2	$\frac{1}{2}(8) + \frac{1}{2}(-2) = 3$
Bet place	5	−2	$\frac{1}{2}(5) + \frac{1}{2}(-2) = 1\frac{1}{2}$
Bet show	3	−2	$\frac{1}{2}(3) + \frac{1}{2}(-2) = \frac{1}{2}$

The optimal strategy, then, is to bet win. Generally, we can say that if the decision maker feels lucky, this is the way to be rational about it.

A third criterion is the logical outcome of the subjectivist approach. This criterion is called the Laplace criterion. We have already noted that the subjectivist maintains that most decision problems under uncertainty from the objectivist point of view can be converted into ones under risk from the subjectivist point of view. However, suppose that the decision maker really knows absolutely nothing about the relevant probabilities. For example, if the writer were faced with the betting decision problem of our example this would be the case because he doesn't even know how to read the relevant racing forms. What does the subjectivist propose to do in this case? After suitable statements about the rarity of this kind of decision problem the subjectivist will generally, if his interlocutor is persistent enough, take refuge in the Laplace criterion. Since this criterion is one of the major attacking points of the objectivist it is important to recognize that the subjectivist maintains that this is a rare case. However, it does occur. What, then, is the Laplace criterion? Simply to convert the decision problem to one under risk by assuming that the probabilities of occurrence of each state of nature are equal. Let us apply it to our example first and then debate it afterwards. Since there are four states of nature in our example we assume that the probability of occurrence of each of them is 1/4 and then calculate the expected value for each strategy:

Don't bet	:	$\frac{1}{4}(0+0+0+0)=0$
Bet win	:	$\frac{1}{4}(8-2-2-2)=\frac{1}{2}$
Bet place	:	$\frac{1}{4}(2+5-2-2)=\frac{3}{4}$
Bet show	:	$\frac{1}{4}(1+2+3-2)=1$

The largest expected value is that of the strategy of betting show so this is the optimal strategy according to the Laplace criterion.

Probably the major reason for the antipathy to this criterion is its close relation to an ancient, and much disputed, philosophical principle called the principle of insufficient reason. According to this principle if there is no reason for something, then it won't happen. Thus, Buridan's ass—of considerable fame in the Middle Ages—would starve to death if exactly equidistant from two equally attractive bales of hay. The reason? Since the ass would have no reason to move to one bale in preference to the other, he couldn't move and, hence, must starve to death. This is an example of quite a number of similar applications of this principle in the past. Suffice it to say that such applications brought the principle into considerable disrepute. Now, of course, the relation of this principle to the Laplace criterion is clear. By the Laplace criterion we argue that there is no reason to assume different probabilities for the various states of nature because we know nothing about them. Since there is no reason for them to be different, then by the principle of insufficient reason we assume they are the same. Despite the ludicrous example of misuse of the princi-

ple of insufficient reason it appears that it has as much justification for its discrete use as any of the other very basic principles used in science—Occam's razor for example. In other words, the relation of the criterion to the principle of insufficient reason does not seem to be adverse to the criterion because, in fact, the principle is a perfectly good one. Physicists dealing with subatomic particles use this principle all the time in the form made famous by Leibniz: two things for which there is no discernible difference are the same thing. Considering that there is rarely a case where the decision maker is really completely ignorant concerning the probabilities of the various states of nature it appears that the Laplace criterion is a reasonable way to fill the gap.

We will consider one more criterion for decision making under uncertainty. This one is due to Leonard Savage and is called the minimization of regret. Savage's criterion is really not a criterion at all. More precisely, it is an alternative way of presenting the payoffs. Savage argues that the chain of events is as follows. The decision maker selects his strategy. Then a specific state of nature occurs and the decision maker receives his payoff. At that time the decision maker will experience a regret due to hindsight: he will reason something like this, "Why on earth didn't I choose that other strategy—in which case I would have made x dollars more?" Savage says that the decision maker should select a strategy to minimize this regret and he proposes to measure the regret by taking the difference between each payoff and the largest payoff for that state of nature. Consider our example. Suppose the horse actually loses. Then if the decision maker had selected the strategy of not betting he will experience no regret but if he selected any other strategy he would experience a regret because he hadn't decided not to bet. Savage suggests that this regret should be measured by the $2 he lost which he could have avoided losing if he had not bet. Suppose the horse actually wins. Then if the decision maker bet win he would experience no regret but if he didn't bet he would experience a regret measured by the $8 he might have won and didn't win. If he bet place he would experience a regret of $6—the difference between what he actually won and what he could have won if he had bet win. Proceeding similarly we can quickly construct the regret matrix:

	States of Nature			
Strategies	*Win*	*Place*	*Show*	*Lose*
Don't bet	8	5	3	0
Bet win	0	7	5	2
Bet place	6	0	5	2
Bet show	7	3	0	2

Now, Savage proposes to apply the Wald criterion to this regret matrix. Since all of these payoffs are measures of regret the decision maker will want to make them as small as possible. The worst that can happen for each of his four strategies, in order, is 8, 7, 6, 7. The minimax regret strategy is, therefore, to bet place.

At this point we will return to an

earlier remark and point out that this is not really a precise application of the Wald criterion to this regret matrix. It is easy to see that the payoff matrix has no equilibrium point so it follows that the optimal strategy should be a mixed strategy. We will not give the procedure for determining such an optimal mixed strategy here.

The Savage criterion seems, introspectively at least, to be a reasonable one. One does experience something akin to Savage's "regret" and it would appear to be good sense to try to minimize it. However, a distinction can be made between "regret" and Savage's suggestion that the Wald criterion should be applied to the regret matrix. One can accept the former without agreeing with the latter. A very strong behavioristic kind of argument can be made for the Savage criterion, however. This is that the criterion is the only one for which it would ever be optimal to use a hedging strategy. Since hedging is a very frequently occurring phenomenon in the business world, it seems that Savage's criterion is more realistic than the other criteria in this regard.

The interesting conclusion of this exemplification of the four major suggested decision criteria under uncertainty is that each of the criteria has selected a different strategy. Specifically, we found

Criterion	*Strategy Selected*
Wald	Don't bet
Hurwicz	Bet win
Laplace	Bet show
Savage	Bet place

This is what we meant when we said earlier that there is an embarrassment of riches in the case of decision criteria for decision making under uncertainty. Each of our four available strategies has been selected by one of the criteria. There isn't even a consensus among the criteria! While arguments can be given for and against each of the suggested criteria the arguments certainly aren't strong enough to lead to a conclusion regarding the general superiority of some one of the criteria over the others. This, at least, is true if we consider the general decision problem.

When we confine our attention to the more specialized subclass of those decision problems for which dollars constitute an adequate measure of payoff there are some powerful arguments in favor of the Laplace criterion. Specifically, the Savage criterion, as we already noted, is not so much a criterion as an alternative way to measure payoffs. When the original payoff matrix is in terms of dollars, Savage's regret measure is nothing other than the definition of economic opportunity costs. Now, economic logic and business commonsense both show that the analysis of any business decision problem should produce the same selection of strategy—whether the problem is analyzed in terms of the original dollar payoffs or in terms of opportunity costs. This means that no criterion should be used which does not produce the same selection of strategy when it is applied to the original payoff matrix and to the opportunity cost matrix (Savage's regret matrix). This eliminates at one fell swoop the Wald, Hurwicz, and Savage criteria. It may seem surprising

that Savage's criterion should be eliminated but this is because he chose to apply the Wald criterion to his regret matrix. The only criterion which will select the same strategy when applied to either of these two equivalent representations of the same business decision problem is the Laplace criterion. Of course, the criterion used for decision making under risk—a generalization of the Laplace criterion—also meets this requirement, as the reader can easily verify for himself.

These, then, are some of the insights which result from the decision theory approach to decision problems. Many questions have been raised and, perhaps, few have been answered but the practicing decision maker can hardly fail to profit from some of the insights developed in this fascinating field. And in the near future we may hope for answers to some more of the questions.

IV GAME THEORY

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE USES OF GAME THEORY IN MANAGEMENT SCIENCE[1]

MARTIN SHUBIK

WHAT IS GAME THEORY?

Game theory is a method for the study of decision-making in situations of conflict. It deals with problems in which the individual decision-maker is not in complete control of the factors influencing the outcome. A general whose forces face the enemy, an industrialist whose products must compete with those of another industrialist, a player in a poker game, duelists, politicians fighting for a nomination, bandits, and bridge players are all involved in struggles which we may classify as game situations.

The essence of a game problem is that it involves individuals with different goals or objectives whose fates are interlocked. There are many examples of decision-making where this is not so. An architect who has been allotted a specified sum of money in order to carry out a given building program or an engineer engaged in redesigning an industrial process in order to cut cost of production are not involved in a game situation. The engineer and architect face direct minimization or maxi-

[1] The work for this paper was done under Office of Naval Research Contract No. N6onr-27009.

Management Science, *October 1955, 2:1, 40–54.*

mization problems in which they are in control of the relevant variables and do not have to contend with anti-engineers or anti-architects who try to destroy their work. The architect may try to maximize certain features of the quality and quantity of building that he can get done for the amount of money at his disposal. The engineer tries to minimize costs for the output of goods required. There may be forces which they do not control, such as the weather; but in most cases some physical law of prediction can be found for estimating the effect of outside influences.

Although it may appear that the Weather wants to rain every time we go on a picnic, unless our pessimism and religion are such, it is not always reasonable to assume that the Weather is a human or super-human agency whose desires are consciously opposed to ours. On the battlefield we may assume that the opposing general is consciously trying to thwart our purposes. The rival firm in business may be actively engaged in taking our customers away from us. In the first case we do not have a game situation; in the other cases we do.

The problem of game theory is more difficult than that of simple maximization. The individual has to work out how to achieve as much as possible, taking into account that there are others whose goals are different and whose actions have an effect on all. A decision-maker in a game faces a cross-purposes maximization problem. He must plan for an optimal return, taking into account the possible actions of his opponents.

THE ELEMENTS OF GAME THEORY

The elements which describe a poker game are the players, money, a pack of cards, a set of rules describing how the games are played, which hands win in any situation that can arise, and what information conditions there are at any stage of the game. The elements which describe the situation of two firms in an advertising campaign are the two sets of individuals in control of the decisions of both firms, the amounts of money available, the information state, the market forecasts of the effect of different types of advertising, and the various laws and physical conditions which delineate which actions are legally or physically possible. The situation in which two opposing field commanders may find themselves can be described in terms of the number of men at their disposal, the amount of equipment, their information and intelligence services; the terrain of the battlefield and weather conditions and their valuation of the importance of various objectives.

All the above examples obviously have a common core. A game is described in terms of the players, or individual decision-makers, the rules of the game, the payoffs or outcomes of the game, the valuations that the players assign to various payoffs, the variables that each player controls, and the information conditions that exist during the game.

These elements, common to all situations of conflict, are the building blocks of game theory. They play the

same role in this theory as do particles and forces in a theory of mechanics. The players and the rules of the game provide a description of the physical situation and the attempt of the players to maximize or to achieve some individual goal provides the motivation or force.

A *player* in a game is an autonomous decision-making unit. A player is not necessarily one person; it may be a group of individuals acting in an organization, a firm, or an army. The feature that distinguishes a player is that it has an objective in the game and operates under its own orders in an attempt to obtain its objective.

Each player is in control of some set of resources. In poker these resources are cards and money; in business corporations they are various assets; in war, men, armaments, and resources. These resources, together with the *rules of the game,* describing how they can be utilized, enable us to work out every alternative that is available to a player. In chess we start with a set of pieces placed on the board in a certain manner; the rules tell us how each piece can be moved. Given that information, we can work out every possible first move that is feasible in a chess game. As we know the initial distribution of the enemy's men and the rules concerning their movement, we can also work out every alternative that he can choose for his first move. In fact, it is theoretically possible to work out the whole game of chess without ever playing it because we could calculate every possible way of playing the game beforehand. Practically, the computation problem is too immense to carry out, but we can imagine a game of chess being played in which each player goes up to the referee, hands him a book containing his complete *strategy* for the game, and then leaves. The referee then works out the game according to these instructions. A strategy for a chess game is a complete set of instructions which states how a player will make every *move* until the end of the game, taking into account all information concerning the enemy's moves. A strategy in war or in business is the same. It is a general plan of action containing instructions as to what to do in every contingency. Thus, the commanding general may tell his subordinates how he wants the attack to begin, then he may tell them what he wants done after the first part of the attack, depending upon what the enemy's actions have been up to that point.

The outcome of a game will depend upon the strategies employed by every player. Let us call the set of possible strategies that the i-th player can use S_i. This is the set of every possible plan of action that the i-th player can have, taking into account his resources, what he can do with them, and also taking into account every possible act by his opponents. Suppose that the i-th player selects a strategy s_i out of all his available strategies S_i. The outcome of the game to him will depend upon what he did and what his opponents did. His *payoff* is a function of the strategies employed by all the players. We can denote the payoff to the i-th player by the *payoff function* $P_i(s_1, s_2, s_3, \cdots, s_n)$. The possible payoffs in chess are win, lose, or draw; in poker they are various sums of money; in business,

profits and growth. In every case each player must have a method of valuation or a utility function which enables him to decide whether or not one payoff is better than another. In business and in games the payoff may be in money and there may be no difficulty in distinguishing between a payoff of $1,000 or one of $100. However, in many cases the payoff can be complicated by other factors. For instance, the payoff arising from following one line of action in battle may result in 1,000 enemy casualties at a cost of 200 men lost; another line of action could result in 5,090 enemy casualties at a cost of 2,000 men lost. It is difficult to say which is preferable.

In general, a player has a valuation scheme by which he can evaluate the worth of any set of *prospects* with which he is confronted. For instance, we assume that a player knows whether or not he would rather make a profit of $1 billion or $10 million. The game in which he is playing may be such that he can never obtain a profit of $1 billion. This amounts to saying that the *prospect* of a profit of $1 billion to a player is not a possible *payoff* in this game.

We may now reformulate the problem of game theory. An N-person game consists of a set of N players, each in control of a set of strategies S_i, $i = 1, 2, \cdots, N$; each player has a payoff function $P_i(s_1, s_2, \cdots s_N)$ which tells him what prospect he receives as his payoff if each player has chosen his strategy s_i. The object of every player is to attempt to obtain a payoff which yields him a prospect of maximal value.

The technical terms described above give us a method whereby we can formalize any sort of situation involving conflict. For the general purposes of the game theorist this is very desirable. However, those of us interested in management science must ask: Can the general scheme be applied to areas of specific interest to us? It turns out that the simplest sort of game we can discuss has several useful applications.

THE TWO-PERSON ZERO-SUM GAME

The two-person zero-sum game is a game in which the amount that one player loses is precisely the amount that the other player wins. Two-person poker, matching pennies, and most other two-person games are of this variety. Competition between two large firms may not be of this type. A price war may damage both of them; collusion may help both of them. However, there are situations in business and war which can be approximated by a zero-sum model.

We can display the relevant features of a two-person zero-sum game by making use of a *payoff matrix*. A whimsical example serves as an illustration. A bootlegger has two possible routes over the border: one is down the highway and the other through the mountains. If he could go down the highway unhampered, he could take a fully loaded truck and make a tidy profit. If there is a light police guard on the main road, he can avoid arrest but will not be able to get his load through and will have to lose the expense of the journey. If there is a heavy police guard on the road, he will be caught and

will be arrested and lose his load. The mountain road is such that he can only take a small load. If it is unguarded, he will have no trouble. If it is lightly or heavily guarded, then he can still get through but will have to bribe the peasants to get him by the police. The police have three alternatives: they can put a heavy guard on the main highway, leaving the mountain route unpatrolled; a heavy watch on the mountain route, leaving the highway unpatrolled; or split their forces and put a light guard on both.

We can display the bootlegger's values for the six possibilities as follows:

	Guard Only Highway	*Guard Both Routes Lightly*	*Guard Only Mountain Road*
Highway	−5	−2	5
Mountain Road	2	1	1

The police's preferences are diametrically opposed to those of the bootlegger; thus, their valuation for any outcome is the negative of his. We call this type of game *strictly determined* because, upon examination of the payoffs, there is a definite optimal choice for the bootlegger which is to take always the mountain road, while there is also an optimal choice for the police which is to guard both roads lightly. Both sides can work out that they can always enforce this compromise on the other but can enforce nothing better. The bootlegger knows that if he chose the highway, the police would try to minimize his gain and could guard it heavily; if he chose the mountains, the police would try to minimize his gain and guard the mountains; but even if they did so, the worst that could happen to him is that he would be able to get a small shipment in after having bribed the peasants. The police argue that if they decided to guard the highway only the bootlegger would use the mountains; if they guarded the mountains only, he would use the highway; if they guarded both lightly, he would use the mountains but would only be able to get a small shipment by them, no matter what he did. We can illustrate these computations by adding a column giving the minimum of each row in the matrix, and a row giving the maximum of each column:

	Strategy of Police			
Strategy of Bootlegger	1	2	3	row minima
1	−5	−2	5	−5
2	2	1	1	①
column maxima	2	①	5	

The column represents the computation done by the bootlegger which tells him what the police could do to him if he chose strategy 1 or 2. The row represents the computation done by the police on the assumption that the

bootlegger would try to maximize against their actions. By choosing the mountains, the bootlegger guarantees for himself the maximum of the minima. By putting a light guard on both roads the police guarantees that the bootlegger can never get more than the minimum of his maxima. But we observe that here the

$$Minimax. = Maximin. = 1.$$

The bootlegger can guarantee a small trade for himself, and the police can guarantee that it stays a small trade, no matter what the other side does. A game which has the property that each side has a strategy which results in the maximin. being equal to the minimax. is said to possess a *saddlepoint.* An economic interpretation can be given to this value. When the bootlegger decides to retire, the market value of his trade should be that amount which yields an income of 1 in the same period as it takes per trip.

Not all games possess saddlepoints and, in those which do not, it is not possible for one side blithely to pick a strategy which guarantees very much. For instance, suppose that there had been a general overhauling of both bootlegging and police techniques. The bootlegger had obtained better trucks, and the police managed to stop bribery in the mountains. The effect of the better trucks is that the bootlegger can get by a light police guard at the cost of some breakages and personal strain. He now can carry a bigger load both on the highway and in the mountains. The effect of the police improvement is that a strong patrol could catch the bootlegger if he were in the mountains. The new payoff matrix is:

	Strategy of Police			
Strategy of Bootlegger	1	2	3	row minima
1	−5	3	6	−5
2	6	3	−5	−5
column maxima	6	3	6	

If the bootlegger persists in sticking to the mountain route, he will be lost; if he keeps to the highway, he will be lost. There is no longer a simple decision to which he can commit himself which will yield him a guaranteed profit, even though his techniques seem to have improved more than those of the police. He still has a profitable trade, however, and there is a way for him to guarantee himself an expected profit by following the actions and precepts of most decision-makers in competitive trades, and that is to take a calculated risk. His problem is to decide how to calculate the risk he should take. He is a prudent man and has no false illusions about the stupidity of the police. He knows, for instance, that the police will never split their forces because this would amount to handing him an income of 3 per period, no matter what happened. The bootlegger wishes to calculate what is the biggest expected income that he can guarantee for himself, regardless of what the police try to do. At the very worst, they could find out his plans and maximize their return, i.e., minimize his return given this information. If he definitely commits himself to one action and plays a

pure strategy, he stands to lose 5. He may, however, decide not to commit himself directly but to choose between his two pure strategies according to some probability weighting. We call the use of such a device, which attaches probability weightings to a set of pure strategies, a *mixed strategy.* By using a mixed strategy he can guarantee an expected profit, even if the police were to find out his strategy.

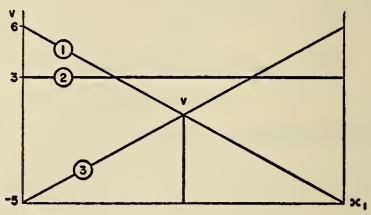
Suppose he decides to take the highway with a probability of x_1 and the mountains with a probability of x_2, where $x_1 + x_2 = 1$. He wishes to pick these numbers in such a manner that he can make his expected return, which we call V, as large as possible under all circumstances. If the police employed their strategy 1 against him, his expected return would be: $-5x_1 + 6x_2$. This must be greater than, or equal to, V. Similarly, for the other strategies, we can write down an inequality. We find that we must solve the following set of equations and inequalities:

$$\begin{aligned} -5x_1 + 6x_2 &\geqq V \\ 3x_1 + 3x_2 &\geqq V \\ 6x_1 - 5x_2 &\geqq V \quad \text{and} \quad x_1 + x_2 = 1. \end{aligned}$$

Similarly, the police wish to make sure that no matter how shrewd the bootlegger becomes they will be able to restrict his expected gains as much as possible. In fact, they can make sure that he never can get more than an expectation of V. The police decide to guard the highway with probability y_1, split forces with probability y_2, and guard the mountains with probability y_3, where $y_1 + y_2 + y_3 = 1$. The police must solve the following set of equations and inequalities:

$$\begin{aligned} -5y_1 + 3y_2 + 6y_3 &\leqq V \\ 6y_1 + 3y_2 - 5y_3 &\leqq V \quad \text{and} \\ y_1 + y_2 + y_3 &= 1. \end{aligned}$$

The general method for the solution of such systems can be found in McKinsey's book;[2] we present a graphical method of this 2 x 3 game.



If the first player uses the probabilities of x_1 and $1 - x_1$ and if the second player uses his first pure strategy, then the expected payoff for the first player is $-5x_1 + 6(1 - x_1) = 6 - 11x_1$; similarly, we get two other expressions if the second player uses his second or third pure strategies. We now draw a diagram with the three lines: $v = 6 - 11x_1$; $v = 3$; and $v = 11x_1 - 6$ represented over the interval $(0, 1)$. For any x_1 chosen by the bootlegger he can guarantee for himself the minimum values of the three lines at x_1; thus, we see here that the optimal choice for x_1 is 1/2. Hence, his mixed strategy uses each of his pure strategies with probability of 1/2 and can be expressed as $(1/2, 1/2)$. It is clear that the police will never use their second strategy; thus, we need only investigate the probability weighting of the police to guard the highway or the mountain road. Their optimal strategy is (1/2,

[2] McKinsey, J. C. C., *Introduction to the Theory of Games,* RAND Corporation, Santa Monica, 1952, Chapters 2 and 3.

0, 1/2). Using these strategies, we can see that the value of the game is $V = 1/2$. Even with better equipment, his business is now worth less than before.

We note that the use of probability in a mixed strategy is a *strategic* use. Any mixture other than an optimal one leaves the player open to damage that he could have avoided, regardless of his opponent's actions.

APPLICATIONS OF TWO-PERSON ZERO-SUM GAMES

1. BUSINESS PROBLEMS

The type of problem to which this theory has a direct application is one which has some of the aspects of a duel. A duel has the property that the goals of the opponents are diametrically opposed. In any market in which the size of the demand is more or less fixed by the government or by habits, the extra customers that one firm can attract must have been lost by another firm. The firms are in pure opposition in such a situation. What one gains, the other loses. An advertising campaign in a market for detergents may be of this nature. A highly simplified example is given below. Charnes and Cooper have written a more detailed paper on this topic.[3]

1a. The Advertising Campaign Two firms, *A* and *B*, each have a million dollars to spend on advertising their products in a certain market area. They can use the media of radio, television, newspapers, magazines, and billboards. For simplicity, we will group these five alternatives into radio, television, and printed media. The marketing research sections of each firm work out the expected effect of any contingency. We will discuss the decision-making at firm *A* only. A payoff matrix of 4 x 4 is drawn up. This contains information on the 16 contingencies that might arise if either firm spent all its money advertising solely by means of radio, television, or printed media, or decided to save the million dollars and not advertise at all. Each entry in the payoff matrix represents the amount of extra revenue above cost estimated under these circumstances (in millions of dollars).

	Radio	*Television*	*Printed Media*	*No Advertising*
Radio	0	− .5	0	2.5
Television	2	0	1.5	5
Printed Media	1	− .5	0	3.5
No Advertising	−2	−4	−3	0

We can see immediately that in this case the alternative of no advertising can be rejected. Any pure strategy, which can be rejected by comparing it with the other pure strategies and finding that there are others which are always better under every circumstance, is a *dominated* strategy and will not enter into a solution. In this simple example, where we have assumed that the firm must put all their money into one advertising medium, we can see

[3] Charnes, A., and Cooper, W. W., "An Example of Constrained Games in Industrial Economics" (abstract), *Econometrica*, 22, October 1954, p. 526.

by inspection that all the other strategies are dominated by television. This game has a saddlepoint at which both firms put their money into television advertising with the net result that they make the same as they would if neither advertised, but neither can risk not advertising.

A more complicated and slightly more realistic example is obtained if we list a series of advertising campaigns involving different integrated programs using more than one medium. Consider each firm to have the choice of three types of campaigns:

	Program 1	*Program 2*	*Program 3*
Program 1	2	4	−2
Program 2	4	2	−2
Program 3	−2	−2	3

The problem for firm A is to find three numbers, x_1, x_2, x_3, such that

$$\begin{aligned} 2x_1 + 4x_2 - 2x_3 &\geqq V \\ 4x_1 + 2x_2 - 2x_3 &\geqq V \\ -2x_1 - 2x_2 + 3x_3 &\geqq V \end{aligned}$$

where the

$$x_i \geqq 0 \quad \text{and} \quad x_1 + x_2 + x_3 = 1$$

The example above has a solution of $x_1 = 1/4$, $x_2 = 1/4$, $x_3 = 1/2$, and $V = 1/2$. Two interpretations can be given to the x_3. They can be regarded as probabilities which should be attached to the decision to adopt any specific program. Or, if it is possible to spend varying sums on the programs (with approximately constant returns), then the x_i give information as to how firm A should split up its advertising budget between the three different programs. It should spend $250,000 on each of programs 1 and 2 and $500,000 on program 3.

The more satisfactory way to treat the advertising problem is as one of a series of games being played every period. Charnes and Cooper suggest this approach in their analysis of Constrained Games in their article noted above.

1b. A Distribution Problem Competition between two refineries sharing a market with relatively fixed demand has been set up and treated as a two-person zero-sum game by G. H. Symonds in his examination of game theory uses in problems of petroleum refining.[4]

2. MILITARY PROBLEMS

Considerable work has been done in the application of the game theory of "duels" to problems of weapons evaluation for tactical weapons. . . . Much of the work in this area is classified. . . .

3. PRODUCTION PROBLEMS INDIRECTLY USING GAME THEORY

3a. Linear Programming Gale, Kuhn, and Tucker[5] have discussed the mathematical analogy that exists between the solution of a linear program and the solution of a two-person zero-sum game. It is always possible to formulate a two-person zero-sum game from a linear program in such a man-

[4] Symonds, G. H., "Applications to Industrial Problems, Including Scheduling and Technological Research" (abstract), *Econometrica*, 22, October 1954, p. 526.

[5] Gale, D., Kuhn, H. W., and Tucker, A. W., "Linear Programming and the Theory of Games," in Koopmans, T. C., ed., *Activity Analysis of Production and Allocation*, John Wiley and Sons, Inc., New York, 1951.

ner that the solution of this game amounts to a solution of the linear program.

3b. The Optimal Assignment Problem Suppose that we have n people available and n jobs to be filled. Suppose further that we have an evaluation a_{ij} which tells us the worth of the i-th person doing the j-th job. The optimal assignment problem [6] concerns itself with the distribution of personnel in a maximal manner. There is a related two-person zero-sum game whose solution gives us a solution of this problem. The difficulty in the application of this comes in the evaluation of the suitability of the attributes of various individuals in the performance of different tasks; although use has been made of this method by the Army.

3c. Statistical Decision Situations in which sampling or gathering extra information costs money, yet cuts down on the possibility of making a wrong decision which may, in itself, be very costly, lead to the formulation of statistical games. In essence, the problem amounts to working out how much one should be willing to pay for information, the value of which will not be known until it is obtained. An example of importance to industry is the design of a decision process to be followed in sequential sampling of a batch of goods where the cost of sampling is high and the loss incurred by sending out a batch with above a certain number of defective items is great.[7]

NON-ZERO-SUM GAMES

Many of the more interesting problems of competition are not zero-sum. The goals of a group of large firms in a market are not necessarily diametrically opposed. There may be "room for all" if instead of fighting among themselves they follow a policy of live and let live. A period of cut-throat competition might hurt all of them. When pure opposition of interests is no longer the case, the computations of the two-person zero-sum game theory no longer apply. No completely satisfactory theory for general N-person games exists at this time. However, the two theories of von Neumann and Morgenstern [8] and of Nash [9] provide much insight into, and useful models of, many situations. It is possible to set up simplified models of some non-zero-sum game situations which merit consideration for application.

1. BUSINESS PROBLEMS

1a. Contract Bidding A firm wishes to bid on some government contracts. It has a certain amount of information concerning the previous behavior of its competitors. Its productive capacity is limited in such a manner that it cannot

[6] von Neumann, J., "A Certain Zero-Sum Two-Person Game Equivalent to the Optimal Assignment Problem," in Kuhn, H. W., and Tucker, A. W., eds., *Contributions to the Theory of Games,* Vol. II, Princeton: Princeton University Press, 1953, pp. 5–12; also unpublished work by Votaw and by Dwyer.

[7] Blackwell, D., and Girschick, M. A., *Theory of Games and Statistical Decisions,* New York: John Wiley and Sons, Inc., 1954.

[8] von Neumann, J., and Morgenstern, O., *Theory of Games and Economic Behavior,* Princeton: Princeton University Press, 1944, Chapter VI.

[9] Nash, J. F., "Non-Cooperative Games," *Annals of Mathematics,* LIV, September 1951, pp. 286–295.

possibly fulfill orders for more than 25 per cent of the contracts. This *N*-person non-zero-sum bidding game [10] can be approximated by a maximization problem which involves picking bids in such a manner that the firm expects to lose 75 per cent of its bids by having named too high a price. This is an example in which it may very easily be to the advantage of the firm involved to go actually to the trouble of randomizing to decide upon certain prices.

1b. The Cost of Price Wars A very important feature of business life which has not received very much stress in economic theory, as it is taught in most institutions, concerns the asset position of a firm and its ability to weather bad times or long fights in its industry. Given information on the asset position of a firm and its major competitors, the expected state of demand for its products and certain other economic data, it is possible to make some basic computations concerning the advisability and profitability of price wars and the introduction of new lines of goods.[11] The major drawback to this work is that few firms have enough information available to make many involved calculations worthwhile at this time.

1c. Checkerboard Land Buying A problem amenable to a certain amount of game analysis may arise when a mining company discovers a tract of land which has promising mineral deposits. The company knows more or less what the land contains and decides that it does not want to develop it immediately. It knows that if it does not buy it, then, sooner or later, the knowledge will leak out and competitors will buy the land. The cost of carrying the land when not using it for anything may involve a tying up of considerable capital. The company needs to work out a maximal strategy of "checkerboard buying" of the land in such a manner that it cuts down on the financing costs by leaving strips unbought, but these strips are unworkable by themselves. It is possible that small competitors step in, buy the strips, and then try to hold up the company when it is ready to develop the land. The company must design its buying strategy in such a manner as to make the holding of these marginal strips as unprofitable as possible to any newcomer. Under certain circumstances it is easy to see that it may pay the company to incur the carrying charges and purchase the whole tract of land. However, if its competitors are financially weak, then they may not be in a position to tie up capital while waiting for the first company to buy them out. . . .

CONCLUSIONS: PROBLEMS AND PROSPECTS IN THE USE OF GAME THEORY METHODS IN MANAGEMENT SCIENCE

Since the war, there has been a great growth in interest in the theory of organization. The size of many modern

[10] Shapley, L. S., "An Example of an Infinite Non-Constant-Sum Game" (unpublished). See also Commander E. D. Stanley, Lieutenant D. P. Honig and Leon Gainen, "Linear Programming in Bid Evaluation" *Naval Research Logistics Quarterly*, Vol. 1. No. 1, 1954, pp. 48–54.

[11] Shubik, M., *Competition and the Theory of Games* (publication in process).

organizations has brought to the surface problems of communication and decision-making of a very different nature to those confronted by smaller groups. A large organization appears to be both quantitatively and qualitatively different from a small one. Information flows and decisions that could be comfortably handled by one "jack-of-all-trades" executive in a small organization or in a dictatorial system, where wastage may be no problem, must be broken down and handled by many specialists. In many cases they may never reach the one-man decision level but are finally acted upon by groups. The need to understand these vital processes of decision-making has impelled us to lay emphasis upon the gathering and study of information, the evaluation of goals, and the role of the individual decision-maker.

The new methods of game theory appear to provide an important approach to many of the problems of decision-making. In this survey of areas of application of game theory some problems which have been completely formulated, solved, and are of immediate practical value have been discussed. . . .

✦✦✦✦✦✦✦✦✦✦✦✦✦ PRACTICAL APPLICATION OF THE THEORY OF GAMES TO COMPLEX MANAGERIAL DECISIONS

SPENCER A. WEART

Frequently, in the management of even the smallest industrial enterprise, complex decisions must be made. Such decisions require that numerous factors must be weighted, one combination of variables equated against all other possible combinations, and a final answer found by some abstruse mental solution of what is really an integral or differential equation. All of this quasi-mathematical process is lumped under the heading of "business judgment," and rarely can the manager trace back the actual steps by which he arrived at his decision in a marginal case, and seldom can he formulate the equations in mathematical terms, despite their real existence. At this point, I hasten to say that managerial decisions never can be purely mathematical, because of the impossibility of assigning exact numbers to

Reprinted from the July–August 1957 issue of The Journal of Industrial Engineering, *8:4 203–209, Official Publication of the American Institute of Industrial Engineers, Inc., 345 East Forty-Seventh Street, New York 17, New York.*

intangibles. But, the mental torture of evaluating the factors can be reduced or eliminated, by analyzing a problem and expressing it in a form that permits a *yes* or *no* decision, instead of a *yes-but* or *no-but.*

The procedure for doing this is spelled out in a complicated mathematical manner in the *Theory of Games.* There is, however, a simple, practical method for applying this interesting theory to complex managerial problems, reducing them to ordinary terms for an easy, reliable, obvious decision.

THEORY OF GAMES

The theory of games is usually considered to have originated in the late 1920's with the mathematician J. Von Neumann, who was the first to show that all games could be expressed in the form of a matrix, that is, a diagram or grid, in which one set of factors is arrayed vertically, another set, horizontally, and values assigned or computed for each intersection. For example:

Should I Carry an Umbrella?

Variable

Choice	*It Rains*	*It Does Not Rain*
Carry	Do Not Get Wet	Do Not Get Wet
Do Not Carry	Get Wet	Do Not Get Wet

In only one of the four possible values is there a possibility of an unfavorable result, which may explain why few people carry umbrellas. But this is only part of the problem, for now the relative weighting of each variable must be considered—what it costs to own and carry an umbrella, and what it costs if I get wet. When mathematical values are assigned to each of the variables, we might get this, for example:

Should I Carry an Umbrella? [1]

Variable

Choice	*It Rains*	*It Does Not Rain*
Carry	$0.50	$0.60
Do Not Carry	$1.25	0

A mathematician will tell you that this matrix figures out [2] at 25 to 2, with a game value of $0.55½, namely, out of every 27 times, on 25 of them I should carry an umbrella, and if I do, I will be $0.55½ better off each time I am right. Now, such a conclusion is really very surprising, and it would

[1] For those who wonder where the values come from the following may suffice: cost of owning a $5.00 umbrella, lost after 10 times carried—$0.50; cost of carrying an umbrella when it does not rain—the above $0.50 plus $0.10 checking charge; cost of getting suit pressed—$1.25.

[2] Those interested in pursuing the mathematics further are referred to "The Compleat Strategyst" by J. D. Williams (McGraw-Hill Book Company—1954) for an elementary but sufficient explanation of how this is done. Those wishing a more complete mathematical discussion are referred to "Introduction to the Theory of Games" by J. C. C. McKinsey (McGraw-Hill Book Company—1952), which has a quite complete bibliography for those wishing to go still further.

appear that the world should teem with umbrella toters. One reason it does not, is that we have not taken into account the chances of it raining while I am outside. If the weather map is such that the chances of rain are 1 to 5; and if my expectation of being outside is 1 to 10, then the odds against being caught outside in the rain are 1 to 50, and the matrix would be:

Should I Carry an Umbrella?

Variable

Choice	*It Rains*	*It Does Not Rain*
Carry	$0.50	$0.60
Do Not Carry [3] ...	$0.02½	0

This time the diagram states unequivocally that I should never carry an umbrella, for whether it rains or not my loss is less if I do not have the umbrella. Since this is more nearly the average situation in real life, few people carry umbrellas in average weather.

In brief, all of the foregoing is a simple application of game theory to what is known as a *2 x 2* game, there being two choices and two variables. If there were three variables the game would be called *2 x 3*, and so on, to *2 x n*. In practice, most business decisions have only two choices, or can be reduced to only two choices—*do* or *do not*, but there are more than two, or *n* variable factors. It so happens that the *2 x n* game lends itself to a very simple graphic and arithmetical solution.[4] Where there are three or more choices, no easy solution is available without resorting to more involved mathematics, or perhaps to a lengthy process of trial and error. Occasionally this may be necessary,[5] but in business it is a rare problem that cannot be reduced to a *yes* or *no* choice, and hence a *2 x n* format.

There are, of course, marginal cases where the answer may not always appear to be the simple *yes* or *no* demanded by the *2 x n* form. The decision cannot be *yes-but*, but it could perhaps be a *yes* handled in a different manner. Should the unprofitable store be continued open or should it be closed? There is another possibility—it might be burned down for the insurance, and with the new working capital thereby gained, re-opened, that is, continued—illegal, perhaps, but nonetheless a possible choice. The answer need not always be a dense black or glaring white, for sometimes a grey answer may be wisest. However, in game theory, each grey can become another choice. If there are more than two possible choices, they can be considered sequentially, but it may be simpler to admit that the game is of a variety higher than *2 x n*.

Assume the question of where a national headquarters office should be located. Should it be in New York City, Chicago, Boston? The choices are three, and the measurable variables can be

[3] Method of applying the odds is explained later.

[4] The procedure is explained later under the caption "An Example."

[5] Games of the *3 x n* type can be solved by a 3-dimensional graph, but *4 x n* and higher types really require the solution of matrices.

many—travel costs to plants, state income taxes, office rent, clerical personnel availability, for example. This situation should not be forced into a $2 \times n$ mold, but should be worked out according to the more complicated rules of $3 \times n$. If there were five candidate cities, a $5 \times n$ diagram would be best. Care is necessary to assure that the problem is not really one of linear programming—the best distribution, for example, of a known limited resource to satisfy a known limited demand. Such problems do not belong to game theory, which is only concerned with the *if* questions—what is the effect *if* this-or-that happens.

A large field of managerial decisions to which no mathematical procedure is applicable is that of human relations. Definite numerical values cannot be assigned to the intangibles of whether or not Jones will make a good sales manager, of whether or not to let the Union have that seniority clause. Here the manager is left to his well-scratched crystal ball. But the fact that game theory does not give an answer to everything need not discourage its use where applicable.

There are innumerable business problems of the $2 \times n$ type, of which the following are representative:

SHOULD A NEW PLANT BE BUILT TO SERVE A CERTAIN MARKET?

There are a number of often conflicting variables: the effect upon present plant overhead through decreased volume—the saving in freight—the reduction or increase in inventories, and in labor or material costs—the cost of newly borrowed funds. While it is not difficult to compute the net gain or loss for a given set of conditions, many contingencies must be faced: If the sales volume does not increase, as hoped? If freight rates rise? If a competitor also builds a plant in that area?

A similar problem is whether two plants should be combined, or one plant divided into two.

SHOULD A CERTAIN PRODUCT BE DISCONTINUED OR ADDED?

Among others, the variables may be: the effects on competing lines, on manufacturing overhead, on sales of nearly similar Company products. Even if difficult to assign exact quantities, the values certainly will range within relatively narrow finite limits. Computations can be made to show what will happen under various possibilities.

A corollary problem would be the raising or lowering of the selling price of one item or line.

SHOULD A COMPANY BE PURCHASED?

Here the variables are many indeed, for frequent unknowns enter into any calculation based upon an assumed static set of conditions. The proper and safe procedure is to set up a game theory diagram, and determine the worst that could happen: the loss of sales, the carrying of an idle plant, the necessity for increased overhead.

A related problem is whether or not a subsidiary company or plant should be sold.

All of the foregoing are essentially *yes* or *no* problems, and the decision, once made, is all too frequently irrevocable. But whereas the number of choices is only 2, the number of varia-

bles is truly *n*—the effect on sales, on competitors, on profits, on operating expenses, and so on, presents a complex of variables, each with its own importance as to weighting, each with its own influence, each contributing its mass, or its little but still not insignificant mite, to the pot which must somehow be cooked up into a decision by the manager. A practical application of game theory can bring order out of this mixture, and array the facts so that an obvious, but otherwise hidden, solution will become visible.

THE PROCEDURE

Before the basic principles of game theory can be applied, the problem must be clearly defined in terms of its variables, reducing each variable to one plain, self-contained question. For example, if a 4½-ounce size of a certain product is discontinued, what will be the effect upon *each* of: (A) sales of the retained 3-ounce size? (B) sales of the retained 6-ounce size? (C) sales of competitors' 4½-ounce sizes? and (D) factory overhead? to name only several. It is not correct procedure to try to consider simultaneously even these few factors.

The proper method is use of the technique of pure research, which holds all but one of the variables constant, and then determines the effect of fluctuations in the one, and only one, variable under study at the moment. A complex problem will become at least to some degree simplified, for the limits of the individual variable are assuredly known, and usually clear. In the example 4½-ounce size problem, the worst that can happen in variable (C) is that competition will get all the previous 4½-ounce sales; the best that can happen is that they will get none of them. Once the maximum and minimum are known, a little thought will quickly lead to a logical, practical figure at some position between the limits. The point here made is that an independent and isolated analysis occurs, disregarding, for example, what might happen to the 6-ounce sales.

Each variable in turn is studied in its own separate, distinct fashion, and the potential gain or loss under the strictly circumscribed set of conditions is established. Values must be in terms of comparable units, the dollar being most suitable. A few variables may be found for which no monetary or definite arithmetical value can be set. Such are few indeed, and where encountered must be left for later consideration, at which time it may be found that they are nongoverning, and can be ignored.

Wherever applicable, odds can be applied before determining the final unit value of the variable. In the umbrella example, the odds of being caught outside in the rain were 1 to 50. That is, the penalty of an unfortunate choice is only 1/50th of the amount first indicated, for the occurrence will happen only once in fifty times. The amount entered in the diagram is properly computed at only this fraction of the otherwise effective value.

After each variable situation has been independently evaluated, all variables are lined up side by side, as in the umbrella example, with still only two choices, but with many additional

variables extending to the right. Here is where game theory really begins to apply, and hence the advisability of restricting the number of choices to two, or at most three, so as to permit a simple mathematical solution.

The final step in pure game theory is determination of the controlling variables, or computation of the correct odds and game value for the specified set of circumstances. The conclusion will be that "the odds are thus-and-so in favor of (against) making thus-and-so much profit." This is a noteworthy answer in itself, and one which many a perplexed manager might well be glad to learn. Even though his final decision might be based upon some (relatively rare) intangible not susceptible of definition in monetary terms, yet this answer would at least have eliminated all other variables, after considering them all simultaneously, each weighted in proper proportion, and giving in one answer the composite result. With this answer the customary mathematical aspects of game theory would terminate, but *the practical application is merely about to start.*

In business, human nature and our economic system being what they are, managers are usually seeking the most favorable results. In game terminology, this would be a *saddle-point* (a pure strategy or sure-fire choice) and high game value; or, if this is impossible, then the most favorable odds, and again a high game value. Where there is no sure choice, game theory terminology calls the answer a *mixed strategy*. Nature being what it is, the initial computations frequently, even usually, are disappointing, and show either or both unfavorable or too low odds, and low game value. The investment of $100,000 in an increased sales force is hardly worth-while if the chances are only 2 out of 7 that $5,000 can thereby be gained. The first computation, which was the final step in game theory, is only a sort of opening gambit.

The variables must be inspected one by one, after an unfavorable answer, to see which are responsible for the unwanted results. When found, each such variable must be searched, once again, to see if it contains within itself some element which alone caused the discouraging results. That increase in the sales force—half the unfavorable expense increase was for a California branch—what if no branch were opened, but the territory merely traveled? A new look with a jaundiced eye may pin-point the part of the program which is causing the bad odds or the low game value—out with it! The part must be sacrificed to retain at least part of the merits of the whole. Conversely, an increase in some expense may tip the balance favorably.

After the search, the isolation of the cause, the re-evaluation of variables, comes the new computation of odds and game value, or if lucky, the finding of the one sure, correct choice. If unlucky, and results are still not favorable enough, then the process must be repeated, and if necessary, repeated again. Finally, the time will come when all possible changes have been exhausted, and the computed results must be accepted. If favorable enough, management's choice is simple—proceed with the plan. If not favorable, the

plan must be abandoned, and lucky the manager who otherwise might have proceeded on pure hunches. A complex problem, with its seemingly infinite variations, has been reduced to simple odds for a given stake, asking the manager only whether the game, at such odds, is worth the gamble. The manager then can hazard the risk in the knowledge that, if wrong this time, in the long run he must win.

It is important that the manager know the principles of game theory and that he help in defining and setting up the problem, but it is not essential that he make the computations himself. In the case of games of $3 \times n$ and higher order the average executive has neither the mathematical inclination nor time to work out the solution. Just as the computation of values for each variable and choice is left to engineers and accountants, so is the solution of the diagram, once set up, best left to these assistants or to the mathematician. There is no objection to the manager figuring the answer himself, but it is really a task for the technician, not the executive. If the latter works through a simple example, he will have sufficient knowledge to exercise top-level judgment.

AN EXAMPLE

An example will be helpful in showing how the procedure may be employed in a practical application. While the variables will differ according to the nature of the problem, the procedure always will be the same.

The example is based upon a problem which vexed a long-established Connecticut manufacturer, John Smith & Sons Company. Smith, among a variety of other household appliances, produces a kerosene cook stove, of simple design, and a meat chopper, both of which are popular in Latin American countries among low-income families. Among these markets, Puerto Rico is not the least in importance. There is, of course, competition from several other American manufacturers, and an increasingly strong potential threat of kerosene cook stove imports from West Germany. There is no foreign threat to the meat chopper line. There is a tariff on imports into the United States and Puerto Rico of cook stoves, and Smith is somewhat hopeful that a quota may be imposed on American imports, but is not too sanguine about it happening soon.

Smith has heard a lot about how various mainland manufacturers have recently set up plants in Puerto Rico, and thinks that perhaps they should build a plant there to supply at least that local market with both the stove and meat chopper; production in Connecticut would be correspondingly decreased. Calculations show that a plant of practical size could sell its output locally, and that a $1.0 million investment would net the company $140,000 annually.[6] Hence the idea seems quite attractive, especially since the company has idle funds of this amount now invested in Government bonds which return only $20,000 per year. The base of the computations as to expected net income is that the Connecticut buildings

[6] All income or loss data in the example are considered as being prior to deductions for Federal Income Tax.

made idle can be sold, and that present conditions as to competition, industry and tariffs will remain unchanged.

But Smith now is faced with the problem of finding, in dollars per year return on investment, whether or not the return is sufficient to warrant the risks of the venture if certain changes do occur. Smith must undertake what amounts to solving a series of simultaneous equations to find one final figure. The variables are these:

1. Effect upon present Connecticut plant if the vacated space is not sold as expected.
2. Effect upon a Puerto Rico plant if the industry-wide volume of kerosene cook stoves decreases. There is some indication that this product may be losing its popularity, being replaced by other types of stoves which Smith does not make nor intend to make.
3. Effect upon Smith Puerto Rican sales of these and other Smith products if a plant is built there.
4. Effect upon Smith's Puerto Rican sales if Smith does not build a plant there, but a competitor does.
5. Effect upon a Puerto Rican plant, if established, if imports of cook stoves from Europe should continue to increase.

While there are other intangible factors, Smith feels that the foregoing are the important points, and that the problems of personnel, local management, etc., can be coped with. Smith already has had experience in starting Latin-American plants and is aware of the intangible problems, which they consider minor.

In terms of game theory, the problem is a *2 x 6* matrix. The *2* represents the choice of *Build* or *Not Build*, since there is no other possible choice. The *6* represents the above five variables, plus the original expected condition. The matrix, or diagram, has as its initial variable column the data representing the assumed, or expected conditions on which the original profitability computations were made:

	Variable
Choice	*Expected Condition*
Build	140 [7]
Not Build	20

The 20 for the *Not Build* choice represents the return that otherwise would accrue from present funds if not invested in the new plant.

Game theory employs the independent analysis of each variable, unrelated to the others at this stage, so each point is considered in turn. We need not be concerned with the computations, which are merely handled by customary accounting and engineering analysis procedures, and will indicate only the answer.

EFFECT OF NOT SELLING PLANT

If the cook stove and meat chopper volume were removed from the Connecticut plant, remaining products must absorb any residual fixed overhead, and this has already been taken into account in the $140,000 profit above computed. But, while the vacated plant space, an isolated building, probably can be sold, the property

[7] Data in all diagrams are shown in terms of thousands of dollars per year.

is not especially desirable and Smith might find that he is unable to sell. Smith then would be required to maintain, insure, and otherwise carry the property indefinitely. This would cost $32,000 annually, so that if not sold, the net annual profit would be only $108,000. The situation as to this variable diagrams as follows:

	Variable
Choice	*Property Not Sold*
Build	108
Not Build	20

EFFECT OF INDUSTRY DECLINE

If the industry volume in kerosene cook stoves should decrease, independently of the effect of imports, operations in a new Puerto Rican plant would in time, of necessity, be curtailed. A reasonable assumption is that within the next decade this might be as much as 50%, in which case the unabsorbed fixed overhead there would amount to an estimated $23,000, thereby reducing the expected profit to $117,000. The diagram as to this situation is then:

	Variable
Choice	*Industry Decline*
Build	117
Not Build	20

EFFECT UPON PUERTO RICO SALES

While Smith now does a substantial volume in both kerosene cook stoves and meat choppers in Puerto Rico, so do several competitors. It appears to be undeniable that a local plant would have a beneficial effect upon Smith sales by diverting a substantial part of competitors' volume to Smith. Also, the demand in the region, especially for meat choppers, is growing, and a local plant is much more likely to acquire a larger share of the increase than a remote mainland plant. If Smith builds, it is highly unlikely that a competitor will do so also. The estimates show that an additional $24,000 of profits can be expected within a few years, or that the profits would be $164,000 instead of $140,000. On the other hand, it is not likely that this additional profit can be secured if the Puerto Rico plant is not built, so that the *Not Build* decision would lose this potential profit of $24,000, less bond income, or a net loss of $4,000. Another diagram thus can be drawn:

	Variable
Choice	*Puerto Rico Sales*
Build	164
Not Build	−4

EFFECT IF COMPETITOR BUILDS

If Smith does not build a Puerto Rican plant, but a competitor does, then the reverse effect of the foregoing can be expected. By the same logic as before, Smith would lose the profit which it now makes on the volume which would be lost to competition. A probable loss of profit, including unabsorbed fixed overhead at the Connecticut plant, is computed at $36,000, less

the bond income, or a net loss of $16,000. The effect of both Smith and a competitor building is ignored, because in practice the first builder will exclude all others. This situation diagrams thus:

	Variable
Choice	*Competitor Builds*
Build	140
Not Build	−16

grammed at three times as much, or $45,000:

	Variable
Choice	*Imports Increase*
Build	−45
Not Build	20

The various factors have now been considered independently, and the next step is to combine them:

	Variable					
Choice	*Expected Conditions*	*Property Not Sold*	*Industry Decline*	*Puerto Rico Sales Increase*	*Competitor Builds*	*Imports Increase*
Build	140	108	117	164	140	−45
Not Build	20	20	20	−4	−16	20

EFFECT OF IMPORTS

The biggest unknown in the entire picture is imports. West Germany is making a drive to re-enter its former Latin-American markets, and is a serious threat to American producers of kerosene cook stoves, even if not of meat choppers. If imports should succeed in getting 80% or more of the cook stove market, as there is a real chance they might, the effect upon the new Puerto Rico plant would be nearly disastrous. If imports did increase to this serious extent, the Puerto Rico operation, instead of showing $140,000 profit, could be expected to show a $15,000 loss. Since best judgment indicates that the probability of a quota being imposed to restrict imports below the 80% of market level are at best 1 to 3, this loss statistically is dia-

The matrix is now complete, and the next step is to compute the odds. There is a simple method for doing this in $2 \times n$ games. First, the diagram is scanned for a *saddle-point*—a value which is the minimum in its row, and the maximum in its column.[8] If there were one, this would show the dominant variable and choice. The best possible choice would be that for the row in which the *saddle-point* lay. For, this would be the least profit that could be made under any circumstance. The only question then would be the managerial decision as to whether this profit was enough. Since there is no such easy answer in this case, we proceed

[8] Strictly speaking, a *saddle-point* is that value which is the largest of the minimums in any row (maxmin) and at the same time is the smallest of the maximums in any column (minmax).

to solve graphically, as in Figure 1. In this figure, for each variable n the proper values are indicated on the *Build* or *Not Build* vertical line, and a line drawn between the points.[9] The highest point on the lowest bounding lines indicates the governing variables. All others can be discarded. The odds and game value are computed on these two variables only—Competitor Builds and Imports Increase:

Build	140	−45
Not Build	−16	20

The method of computation consists of merely subtracting the right-hand figure from the left:

Build	185
Not Build	−36

Disregard the sign, reverse the figures, and the odds are as follows:

Build	36
Not Build	185

That is, out of every 221 chances, only 36 times should Smith build, or, the odds are 36 to 185 against building successfully. The game value is computed by applying these odds to either variable (as a check, the answer must be the same for each):

$$\frac{(36 \times 140) + (185 \times -16)}{36 + 185} = 9 +$$

Smith has only 36 chances out of 221 of being sure of making at least $9,000 per year out of a $1.0 million investment if the plant is built—quite a different story from the $140,000 return under the expected conditions. Should the project be abandoned? Not at all. This is only the first trial run.

The next step is to inspect the diagram, to find the reasons underlying the unhappy results. It is at once plain that the only losses in the *Not Build* row are caused by the failure to have a plant of any kind in Puerto Rico. That is, a plant there would, no doubt, keep out a competitor, forestalling any diversion of sales to him, and also have a beneficial effect on the sales of all Smith products. It is also evident that the only loss in the *Build* row arises from the threat of imports of kerosene cook stoves.

The solution is clear—have a plant in Puerto Rico, but do not make cook stoves there, only meat choppers. Smith then starts a new round of computations along preceding lines, to determine the feasibility of having a smaller Puerto Rican plant manufacturing meat choppers only. This is found to be possible from an operating point of view, with a smaller investment of $750,000 and a smaller $96,000 profit under expected conditions. The resulting diagram omits those variables exclusively concerned with kerosene cook stoves:

[9] Those mathematically inclined will at once recognize this as being merely the plotting of the equations $y = 140x + 20\ (1 - x)$, etc., in which x represents the *Build* choice, and hence $(1 - x)$ the *Not Build* choice and y the pay-off for this variable. The range of x is, of course, from 0 to 1.

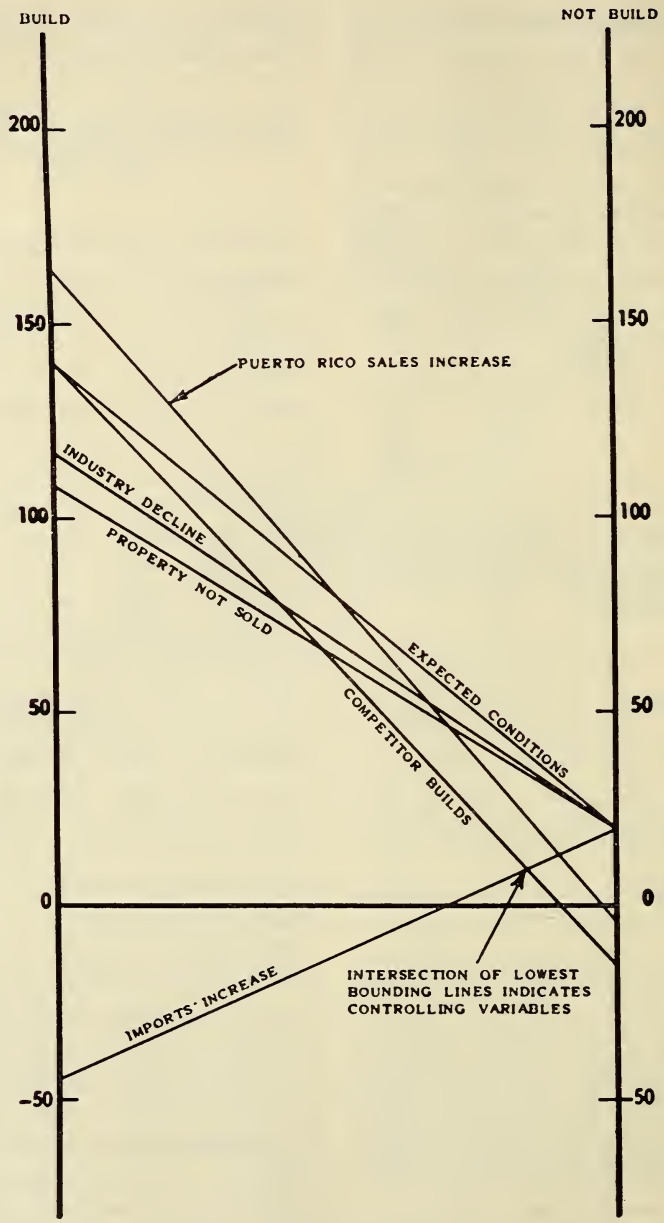


FIGURE 1

GRAPHIC SOLUTION OF 2 × N GAME TO FIND CONTROLLING VARIABLES.

	Variable			
Choice	*Expected Conditions*	*Property Not Sold*	*Puerto Rico Sales Increase*	*Competitor Builds*
Build . . .	96	75	113	96
Not Build	15	15	15	−2

Since, for every variable, the *Build* row is well in excess of the *Not Build* row, the choice is obvious, and since the minimum value in this row, $75,000, represents a 10% return on investment, Smith will build the plant. The decision is made with confidence, for it was based on true business judgment, assisted by logic and mathematics, and not dependent solely on hunches.

C

Tools for Coping with Lack of Information

I SAMPLING

AN INTRODUCTION TO SAMPLING

JAMES H. LORIE AND HARRY V. ROBERTS

. . . The word "sample" is probably already familiar to almost everyone, but just to make sure that the following discussion is understandable, a word about the nature of samples will be injected at this point. The investigator of any problem is inevitably concerned about some characteristic of some group (the group need not be a group of persons —it could be horses, telephone poles, stores, rivets, or shirts). . . . [A] toothpaste manufacturer . . . for example, . . . [is] probably interested in the attitude toward the new flavor of his tooth paste of all persons in the United States who might conceivably buy the tooth paste. If these attitudes were completely known, the "whole truth" would be known as far as this particular project is concerned.

All potential buyers of tooth paste in

From Basic Methods of Marketing Research, *by James H. Lorie and Harry V. Roberts,*

the United States—from whom the information on attitudes might be obtained—can be considered to be a "universe" or "population." (These two terms are interchangeable.) More generally, any complete group in which an investigator is interested is called a universe or population.

Sometimes, as in the decennial count of the population of the United States, a complete enumeration or "census" is taken. Typically, however, it is impossible to examine the entire population; the investigator must be satisfied with something less. For the tooth-paste manufacturer this something less would consist of the knowledge of some fraction of the total number of persons whose preferences were of interest. This fraction of the total number of individuals would be called a "sample." In general, a sample is a part of a universe, regardless of how that part is chosen. Since something less than the "whole truth" is known when samples are the source of information, the possibility of error is introduced. The sampling problem consists in selecting the sample in such a way that this error is controlled to make it as small as possible for a given dollar expenditure.

Sampling can sometimes be avoided and a census taken. It is interesting to note, however, that the sampling method is extremely pervasive. The two most important and obvious reasons for its pervasiveness are the savings in cost and in time. It should also be realized, however, that sampling assists in improving the reliability of the process of communication or observation. It is much cheaper and easier to select, train, and supervise 50 field interviewers who conduct 3,000 interviews than it is to select, train, and supervise 150,000 interviewers who conduct a complete census of a human population. Thus, although sampling introduces "sampling error" which is absent when a census is taken, the *total* error in the research project—that is, errors arising from failure to achieve perfect communication, errors in recording, errors arising in tabulating, etc., as well as error arising from sampling—may actually be less when samples are used. For example, the relatively careful training and supervision of field-workers which the process of sampling makes possible may so reduce errors in the information secured from each respondent—*i.e.*, "nonsampling error"—that the introduction of sampling error is more than offset. One very interesting example of the superior results which can be obtained from the sampling process has been reported by P. C. Mahalanobis.[1] In this example, a sample costing much less than an attempted census of the same population was much more accurate.[2] The reason was simply that the census was such a staggering job that its results were

[1] P. C. Mahalanobis, "A Sample Survey of the Acreage under Jute in Bengal," *Sankhya*, Vol. 4, pp. 511–530, 1939.

[2] The reader may wonder why, in view of the advantages of sampling, the entire population of the United States is enumerated completely every ten years by the U.S. Bureau of the Census. Aside from the fact that the Constitution requires that this be done, perhaps the most important reason is that information is required for very small groups of the population—such as small towns, small areas of cities, etc.—as well as the whole country. Even so, however, about half the questions on the 1950 census were asked only on a sample basis.

very inaccurate. These three reasons, then, savings of time, savings of money, and the possible reduction of nonsampling errors, provide the rational basis for the use of sampling.

THE NATURE OF SAMPLING THEORY

Most people would probably accept the assertion that a sample can provide useful information about the population from which it is drawn. There have been apparent failures of sampling methods such as the 1948 election predictions, where rightly or wrongly some of the blame has been popularly fixed on sampling, but the results of sampling surveys have been surprisingly widely accepted. In part, this acceptance reflects a tendency to believe "figures" without realizing clearly that they have been obtained from samples, but it also stems from the respectability which has come to be associated with sampling. Nonetheless, sampling theory and common sense are not always in accord, and when one is personally affected adversely by the results of a sampling survey, it is easy to find strong intuitive reasons for disbelieving the survey. For example, the point system for establishing priority of discharge from the army after World War II was based (democratically) on a sampling of soldier opinion. Yet one of the favorite comments of those who felt themselves unfairly treated by the point system was, "I'd like to find one of the guys they talked to." Fred Allen has turned skepticism into humor in his comments on Hooperatings a sampling measure of the audience of radio programs: "The 'Hooperating' is a so-called service that allegedly tells you approximately how many listeners the average radio program theoretically has. It's like taking a bite of one roll and telling you how many poppy seeds there are in the country." [3] Allen's jibes came at a time when these ratings indicated that his audience was decreasing.

Suspicion, disbelief, or uneasiness about sampling is also encountered among business executives. In part this feeling arises from findings which contradict well-established beliefs. But this is not all: people responsible for important decisions based on the information obtained from a sampling survey are not always satisfied by glib reassurances that the sample is "broadly representative" or that it provides a "miniature cross section" of the population. . . .

Unless the reader is already familiar with the logical structure of sampling theory, it might be helpful at this point for him to ask himself what reasons he can advance to explain why samples can be expected to "work." Aside from a possible citation of instances in which sampling *has* worked, he will probably have trouble in thinking of convincing explanations. The explanations are relatively simple, yet they probably would not occur to anyone without some knowledge of statistical theory.

Consider an example that dramatizes this point. In the depression years of the 1930's many estimates were made of the extent of unemployment, and many different methods underlay these estimates. The differences were huge—in 1932 there was a difference of several million people between the highest

[3] Quoted in *Newsweek*, Jan. 17, 1949, p. 48.

and lowest estimate. Furthermore, there was no way of being reasonably sure which of these estimates came the closest to the truth. Today, by contrast, there is one widely accepted estimate of employment and unemployment. This estimate is based on a sample of approximately 25,000 to 30,000 families taken monthly by the U.S. Bureau of the Census, which gives results within tolerances for sampling error that can be (and have been) computed and published. Now, suppose in a particular month the unemployment figure is reported by the Census Bureau as 3,500,000. Someone claims this is wrong —that there are really 7,000,000 unemployed. What argument can be advanced—other than arguments of authority—for the contention that 3,500,000 is likely to be much closer than 7,000,000 to the correct figure? [4]

Fortunately, there are two distinct justifications for sampling and the results based on sampling. The first is purely empirical. People have tried sampling experiments with known populations or universes, such as red and white beads in a box, and have found that the results conform with what would be expected from the abstract considerations of sampling theory. Suppose there were 500 red beads and 500 white beads in the box. If the beads were well mixed and 100 beads taken from the top of the box, we usually would find—as we would expect from sampling theory—between 40 and 60 red beads. Perhaps a good way to acquire a "feel" for sampling is to play a relatively simple chance game, such as matching pennies, and to observe the connection between what actually happens and what one would expect from sampling theory. One of the most important lessons of such a game—or of drawing beads from a box—is that untypical results occasionally occur. These "aberrations" might be written off by those ignorant of statistics as good (or bad) luck which violates the law of averages. Statistical theory makes possible, however, not only the prediction of what usually will be found but also the anticipation of such unusual occurrences and the making of precise statements about their frequency (though not, unfortunately, their exact *time*) of occurrence.

The second justification of sampling is a logical one. While a rigorous explanation would necessarily be mathematical, the main steps can be described in intuitive terms with only occasional recourse to symbols. . . .

As has been said, whenever samples are used, there is the possibility of error. If 100 beads were drawn from the box previously referred to (which contains 500 red beads and 500 white beads), it would be a relatively unusual occurrence to obtain *exactly* 50 red and 50 white beads. It is even possible, though exceedingly improbable, that all 100 beads would be red. Hence samples typically do not mirror *perfectly* the characteristics of the population from which they have been drawn. It was said earlier, however,

[4] There have been criticisms of the census estimates. These criticisms center on the way in which one *defines* employment and unemployment. This is not strictly a statistical problem; the *sampling* methods used are apparently accepted by most critics. See *Fortune*, September, 1949, "Those Unemployment Figures," p. 76.

that ordinarily between 40 and 60 red beads will be found in any sample of 100. This assertion was made on the basis of a knowledge of statistical theory. More completely and precisely, the following assertion would have been made: The number of red beads in a sample of 100 beads from this box will fall somewhere in the range of 40 to 60. While this prediction may turn out to be incorrect for any particular sample, the risk of error is calculated (by methods not shown here) to be approximately 5 per cent. That is, in 5 per cent of all samples for which predictions of this kind are made, the predictions will turn out to be wrong. The interesting part of this assertion is that a numerical value (probability) has been attached to the risk of error—the risk that the method which enables the prediction to be made will lead to the wrong answer. . . . Whenever it is possible to so evaluate the risk of error, the samples obtained are called "probability samples." Roughly speaking, probability samples are samples for which the risks of error are measurable. Other types of samples (nonprobability samples) are of great importance and are, in fact, even more widely used than probability samples. . . . But samples of both types can probably best be understood if probability samples are explained first. . . .

SOME FUNDAMENTALS OF SAMPLING THEORY

INTRODUCTION

We are concerned with sketching the outlines of sampling theory. Our objective in doing this is neither to treat sampling theory exhaustively nor even to plunge very deeply into it, but rather to indicate its main steps in a nontechnical way so that the practical applications . . . can be better understood. Above all, we hope to indicate briefly how sampling theory can give confidence in inferences drawn from samples because of knowledge of the procedure by which the sample was selected rather than because the sample results conform to the preconceived views of the investigator. In the following pages, then, we consider the sequence of main definitions and ideas that are needed in order to accomplish this understanding of sampling theory.

POPULATIONS AND SAMPLES

The population or universe may be thought of as the "whole truth," or all possible measurements relevant to some particular question. In actual practice the whole truth is usually unknown. If it were known, neither samples nor marketing research would be needed. Two possible universes are (1) all housewives in the United States and (2) all Dodger fans in Brooklyn.

These examples, however, are slightly misleading. In these cases the universe or population has been "personalized" by thinking of it as consisting of people, families, institutions, or the like. Often it is preferable to think of *measurements* made on people, institutions, etc., as being the universe. Thus, for example, one universe would consist of the most recent purchases of flour by housewives in the United States; another, of the ages of these housewives; another of their attitudes toward radio commercials. In each case

the same people—American housewives—would be involved. In each case, however, the population would be different. It may seem that this distinction between people and measurements made on them is artificial. For ordinary speech it is artificial and throughout our discussion we shall use such short cuts as "the universe of housewives." Nonetheless, the possibility of a distinction should be seen at this point. For by seeing the possibility of making such a distinction, it is easy to understand a problem of great practical importance in sampling. That is, the sampling methods best suited for studying flour purchases of a certain group of people may be different from those best suited for studying attitudes of this same group toward radio commercials or for finding out about some other characteristic of the persons involved.

A mean [average], a proportion, a median, a standard deviation, or any other summary measure of a characteristic of members of a population is called a "parameter." (The analogous summary measure for a sample is a "statistic.") For example, 0.04, or 4 per cent, of the labor force might be unemployed; the median income of families in Chicago might be $4500; or the mean age of people attending movies might be thirteen years. A useful convention in statistical writing is to use Greek letters to symbolize parameters. Two of the most frequently used parameters are the mean of a population, represented by the Greek letter mu or μ, and the standard deviation of a population,[5] designated by the Greek letter sigma or σ. Many statisticians use this apparently esoteric symbolism to avoid a serious confusion, which is explained in a moment.

A sample is any part of a population. If 100 beads are drawn from the box containing 500 red beads and 500 white beads, the 100 beads comprise a sample from the population of 1000 beads. If 100 housewives are interviewed as to their attitudes toward radio commercials, these 100 housewives would be a sample from the population of all housewives. The definition of "sample" does not specify that the sample be drawn in any particular manner. An understanding of sampling theory will, however, be facilitated by thinking of the many possible different ways of selecting samples as falling into one of two major categories, probability and nonprobability samples. In the sketch of sampling theory in this chapter we consider only the most familiar

[5] The following brief comments are added for those who need review on the slightly elusive concept of "standard deviation." The standard deviation is one of a large number of summary measures known as "measures of variation" or "measures of dispersion," since they measure the variability or dispersion encountered in a set of data. One other common measure of variation, for example, is the range, which is simply the difference between the largest and the smallest observations in a group of observations. The standard deviation is defined as follows: (1) Obtain the deviation of each observation in a group of observations from the mean of all the observations in the group. (2) Square each deviation, and add the results. (3) Divide by the number of observations. (4) Take the square root of what remains. (Actually, for computational purposes, there are convenient short cuts which can be found in any elementary statistics book.)

type of probability sample, the simple random sample. . . .

Just as a mean, a proportion, a median, or a standard deviation of a population is called a "parameter," so a mean, a proportion, a median, or a standard deviation computed from a sample is called a "statistic." Sample statistics are used in drawing inferences about population parameters. In drawing such inferences, of course, it is necessary to remember that a sample will almost never mirror perfectly the population from which it is drawn and that statistics of a sample will almost never equal the population parameters. To emphasize this fact, many statisticians follow the practice of using Roman letters to symbolize sample statistics. The sample mean, for example, is called $\bar{x}$ (pronounced "$\bar{x}$ bar") while the sample standard deviation is called s. It is essential to distinguish $\bar{x}$ and s from μ and σ, the symbols for the mean and standard deviation of a population. We use all these symbols in the discussion of sampling theory, and any confusion between them will lead to confusion in the understanding of the basic ideas of the theory.

PROBABILITY AND RANDOMNESS

The two closely related concepts of randomness and probability supply the foundation of the theory of statistical inference and, in particular, of the theory of sampling. Probability has a nontechnical, although imprecise, meaning which is familiar to everyone. Everyone is familiar with sentences such as, "His chances of winning are pretty small," "It's pretty likely that we'll have rain before tomorrow," or "You're probably right." In each of these examples the general idea of uncertainty is involved, and by use of words such as "chances," "likely," or "probably" this fact is acknowledged. All these terms are closely related to the nontechnical idea of probability.

Probability has a much more precise and subtle meaning in statistical inference. Without delving into the subtleties of this meaning, we shall first make a statement that conveys a part of the meaning and then modify this statement in order to convey a fuller understanding. The simplest definition of statistical probability (which we amend shortly) states that the probability of a particular event occurring under given conditions is equal to the relative frequency of its occurrence under those conditions. If 950 out of each 1,000 live-born babies reach the age of one year, the probability of survival to the age of one year would be 0.95 (*i.e.*, 950 divided by 1,000). If a coin seems to come up heads in about two-fifths of all tosses, the probability of a head would be said to be 0.40. Relative frequency can be as small as 0 if a particular outcome never occurs and as high as 1 if it always occurs. In most discussions it is customary to speak of an outcome with a probability of 0 as being "impossible" (since it never happens) and an outcome with a probability of 1 as being "certain" (since it always happens).

This first approach to an understanding of probability, then, equates probability with the familiar term "relative frequency." If one stopped with this definition, however, there would be trouble. For example, half the people

in the United States are men, and therefore the probability that the mayor of Chicago is a man would seem to be one-half. This is of course absurd. A person is or is not a man, and the probability is certainly not one-half. People, or beads, or opinions are what they are and cannot be "probably this or probably that." Probability applies only to predictions about these immutable facts. To obtain a more adequate idea of probability, the closely related idea of randomness must be explored.

A concise, clear, and correct definition of randomness is beyond our present needs. We shall, however, attempt to get an intuitive grasp of the idea by proceeding with an example. Imagine a perfectly balanced roulette wheel with just 10 numbers, the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. Someone spins the wheel and waits for it to stop. When it stops, a pointer points at one of the 10 possible numbers. The number is recorded and the wheel is spun again. Each time the wheel is spun, the number designated by the pointer is recorded. After a series of spins, the following sequence of outcomes is obtained:

8092 9572 9931 3826 9066 0003
2184 4003 3574 3408 7331

A person would probably feel intuitively that each of these 10 digits would tend to have a relative frequency of about 1/10 if the wheel were spun repeatedly. If, before each spin of the wheel, one were to predict that the digit 7 would turn up, he would be correct about 1/10 of the time. And there is nothing that can be done to make the proportion of correct predictions significantly greater than 10 per cent. The reason is that while 7 will turn up about 1/10 of the time, it is impossible to know *when* it will turn up. For the sequence of observed digits has no apparent pattern or regularity. Or, if evidence of regularity is observed in a part of the observed sequence, there is no assurance that this pattern will be of any use in increasing the number of successful predictions.

It is possible to think of the sequence of digits produced in this manner as a *sequence of random observations.* We have not really defined how this sequence was produced; we have simply described a mechanical process (spinning a roulette wheel) and assumed that this process would in fact lead to a sequence that lacks any persistent pattern in the order of digits produced.

A random sequence, then, may be thought of as a sequence of observations like that produced by some mechanical device similar to one described here. The mechanical process is not defined precisely; it must be understood intuitively. A process (mechanical or otherwise) is random if it gives rise to a sequence of observations having no regular discernible pattern.

In view of the vagueness of the definition of a random process, it is especially desirable to have some way of knowing whether any particular process can be regarded as random. The only way to decide whether a process can be regarded as random is to judge by the appearance of the sequence of observations produced by the process. There are tests by which one can arrive at a judgment as to whether the

process is random. Essentially these tests can be thought of as tests for evidence of persistence or regularity of pattern in the sequence of observations.

Long sequences of digits, similar to the short one given above, have been produced (though not with roulette wheels) and tested for randomness. These sequences are known as "tables of random digits." [6] Tables of random digits, in turn, are immensely useful in practical sampling applications, as will be seen later.

We are now in a position to complete the definition of probability: "The probability of an event is approximated by its relative frequency in a very long sequence of random observations." In the above example, we might therefore state that the probability of the outcome 7 in spinning the roulette wheel is approximated by its relative frequency in a very long sequence of such spins. We have assumed that this relative frequency would be about 1/10.

The reason for probing this far into the definition of probability must now be explained. If, in the selection of a sample, one is able to assign a known positive (nonzero) probability of selection to every member of a population, statistical theory can be used to make rigorous inferences about the population from the information in the sample. There is no other known way in which rigorous inferences can be drawn. When the process of selection is such that a known positive probability of selection can be assigned to every element in the population the resulting sample is a probability sample. The only way in which a known positive probability of selection can be assured, as we have seen from our more complete definition of probability, is by a process of random selection, since probability is neither defined nor measured in the absence of randomness. The simplest process of random selection is one which assigns an *equal* probability on each draw to each member of the population. Such a process of selection is called "simple random sampling," and the samples that result from such a process, "simple random samples." . . . We discuss only simple random samples.

The next task is to see specifically how a simple random sample would be obtained. To obtain such a sample, a table of random digits may be used. To illustrate the use of such a table we shall repeat the sequence of digits produced by the roulette wheel (which, in principle, would be capable of producing a whole book of random digits).

8092 9572 9931 3826 9066 0003
2184 4003 3574 3408 7331

This sequence of digits resembles a single line of digits in a table of random digits which, in turn, consists of several pages filled with digits in this fashion. Suppose that the particular line of digits reproduced above is one line in such a table.

The problem is to draw a random sample of charge-account customers in a large department store. This depart-

[6] One example of such a table is given in George W. Snedecor, *Statistical Methods*, Chap. 1, Iowa State College Press, Ames, Iowa, 1946.

ment store has 10,000 such customers. A list of these customers is made and the customers are assigned consecutive numbers as follows: 0000, 0001, 0002, 0003, ... , 9997, 9998, 9999. The next step is to refer to a table of random digits and proceed consecutively from a starting point just as in reading a book. If the starting point had been the beginning of the line of random digits shown above, the 8092d customer would have been selected, since the four digits 8, 0, 9, and 2 were the first four encountered. The next four digits are 9, 5, 7, and 2; hence the next customer selected in the sample would have been the 9572d. One proceeds in this manner until the desired number of customers, say 225, has been obtained.[7]

[7] For most sampling purposes, if the same number were to turn up twice, the second appearance would be ignored. Technically, this method of sampling is described as "sampling without replacement." One other possible complication is also worth mentioning since it causes some confusion. If there had been (say) 9,214 customers instead of 10,000, the procedure would have been exactly the same with one exception: all numbers greater than 9,213 would have been ignored if they had been encountered in the table. Hence 9,572 would have been skipped.

Before going on it is worth while to refute two fallacies about random sampling that are rather widespread. First, the idea of randomness is not the same as its everyday connotation of "haphazardness." The reason that tables of random digits are accepted is that they have been subjected to tests for randomness. By contrast, haphazard sampling processes, in which the sampler tries to be "aimless," seldom succeed in approaching true randomness.

Second, a random sample is not necessarily a representative sample or a true cross section. It has already been seen that it would be possible to draw by a random process 100 red beads from a box that contained 500 red beads and 500 white ones. The probability of this happening in random sampling, fortunately, would be very small. But if such an occurrence did come about in random sampling, the 100 red beads would be a random, although a most unrepresentative, sample. Whenever the term "representative" is used in practice, it is necessary to examine carefully what is meant, since, unless the characteristics of the population are known exactly, it is impossible to assure the selection of a perfectly representative sample.

II STATISTICAL INFERENCE

✦✦✦✦✦✦✦✦✦✦✦✦✦ ANALYSIS OF SAMPLE DATA

ROBERT FERBER

All sampling operations have one or both of two ultimate objectives in mind. One objective is that of obtaining as accurate information as possible of the value of certain population characteristics (parameters) from the sample data; this is the problem of *estimation*. The characteristic being sought may be a single figure, such as the average soap purchase per family in the United States, or it may be an entire distribution, such as the average soap purchase per United States family by income level. It may deal with only one specific subject, or it may comprise a whole range of subjects, as do some consumer panels. It may cover only one period of time, or it may cover several periods of time. In other words, there is almost no limit to the purposes to which samples are, and have been, put.

Alternatively, the purpose of a sampling operation may be to test some theory about the composition of the population, in which case the acquisition of sample data is not the primary aim, but is simply the means to a further end; this is the problem of *testing hypotheses*. An example of this latter type is the situation where, knowing the 1944 regional purchase pattern of coffee *X*, the research director is eager to determine whether there has been a significant change in the purchase pattern of that commodity by 1946, and thereupon samples the population to determine the 1946 regional purchase pattern. With the aid of this sample-determined 1946 regional purchase pattern, the significance of the change from the 1944 regional purchase distribution is then determined. Of course, the purpose of the sample may be twofold; the research director in the above example may be just as anxious to ascertain the 1946 regional purchase pattern as he is to determine the significance of the difference between the 1944 and 1946 purchase distributions.[1]

Estimation. Only by the purest coincidence will a sample ever provide a perfect representation of the population. This is an unavoidable conse-

[1] The former might be used to estimate future sales and set sales quotas; the latter to test the effectiveness of advertising campaigns.

quence of the erratic variations introduced by the sampling process itself, variations that cause the sample value to deviate from the true population value by a margin indicative of the deflecting effects of random sampling influences. Thus, if the average height of all United States males is 68.8 inches, a sample of several thousand men may have a mean [average] value of 68.6 inches or of 68.9 inches, but only by coincidence will it have the same mean value as that of the population. Therefore, any estimate of the true population value based on sample data must contain some allowance for such random sampling variations. In other words, the primary function of a sample in estimation problems is not to yield a *point* estimate of the population value but to provide a *range* of values within which the true value is thought to lie.

As a consequence of the development of the theory of probability, this allowance for, or range of, random variations can be measured statistically. If a great many large fixed-size samples are taken from the same population, it is known that the mean values of the samples will tend to be normally distributed around the mean value of the population, so that, for example, approximately 68.27 per cent of the sample means will be contained within the interval of the population mean plus and minus its standard deviation. Consequently, by working backward and estimating the standard deviation of the population characteristic from sample data, it is possible to estimate the range within which a sample mean is likely to deviate from the true population mean. Thus, if 68 per cent of the sample means are known to lie within plus and minus 1 standard deviation of the true mean, then there is a 0.68 probability that the mean of any *one* sample is within this interval.[2] Conversely, if an infinite number of samples were drawn from this population, we would be correct 68 per cent of the time if we stated, in each case, that the population mean was within the interval of the sample mean plus and minus 1 "standard error" of that mean. The standard error is the estimated value of the (unknown) standard deviation of the sample means in the population, i.e. estimated from the sample data. Now, if only one sample has been taken, which is the usual case in practice, we would have a 0.68 probability of being correct if we were to state that the true mean lies within the interval of 1 standard error of the sample mean. This interval is known as a *confidence interval,* the associated probability being known as the *confi-*

[2] Note that the theory is couched in terms of the probable deviation of the sample mean from the population mean. The reason for this is that the true population value in any problem is always fixed, though unknown. Therefore, one cannot speak of the *probable* distribution of a population mean about a sample mean, as there is no element of probability as to what the population value is. The element of probability enters into the determination of how accurately it is possible to estimate the population mean from sample data. The true average height of United States males may be 5 feet 8 inches; this, though unknown, is a definite fixed value. But the average height of United States males *as estimated from a sample* will not be fixed but will vary from sample to sample. It is this variation of the different sample means about the true population mean that the above theory seeks to measure.

dence coefficient. Hence, an interval having a 0.68 confidence coefficient means that the true population mean will lie within the interval of 1 standard error of the sample mean in 68 samples out of 100 (all of the same size and drawn from the same population).

As noted above, the sample mean plus and minus 1 standard error provides us with a 0.68 confidence coefficient. If a higher degree of certainty is desired, a larger confidence interval would have to be employed, say, the sample mean plus and minus 2, or 3, standard errors, in which case the confidence coefficients would increase to 0.955 and to 0.997, respectively.

The numerical value of these standard errors is computed by means of the standard-error formulas. The probable range within which the true population value is likely to lie, the confidence region, or the confidence interval, is obtained as a multiple of these standard errors. It is this computed range that, together with the average, or aggregate, sample estimate, furnishes the final estimate of the population value. It should be emphasized, however, that the sample estimate [3] by itself is not a satisfactory estimate of the population value, as the mathematical probability of a sample estimate coinciding with the true (unknown) population value is approximately zero in most of the usual populations; it merely serves as the reference point for the construction of the final estimate of the confidence region.

The following example illustrates this point. To estimate the average value in a population as 50 units simply because the average value of the sample comes out to be 50, without specifying the value of the standard error, is meaningless, for one has no idea of the distortion introduced into the estimate by erratic sampling variations. If the standard error is computed to be 1 unit, then one can be fairly sure that the true population value is about 50.[4] On the other hand, if the same sample value has a standard error of 15 units, very little reliability can be placed in the sample figure of 50, as the high value of its standard error indicates that the confidence interval for the true population value is between 20 and 80—using the sample mean plus and minus 2 standard errors to indicate the range within which erratic sample variations might cause the sample mean to deviate from the true figure.

The theory of constructing confidence regions presents many separate problems of its own, but it is inherently linked to the problem of statistical estimation, for unless confidence regions are specified, estimates based on samples are practically valueless. As will be pointed out later, the preferability of different sampling techniques rests almost exclusively on a comparison of the relative size of the confidence regions they may be expected to produce, and *the ultimate objective of all sam-*

[3] By which is meant the central sample value, or statistic. The wording is rather ambiguous here, for the *sample estimate* is an estimate of a *population* value, not of anything in the sample, as the term may imply. Furthermore, it is only a preliminary estimate, as a particular sample estimate will almost never coincide with the actual population value.

[4] Assuming absence of bias in the sample.

pling research is to develop techniques that will either yield the smallest confidence region at a given cost or a given confidence region at the most economical cost. . . .

Testing Hypotheses. The validity of certain inferences about the nature or composition of the population is confirmed or disproved on the basis of statistical significance tests on the sample data. The criterion for these tests is to determine whether the observed difference might have occurred as a result of random sampling variations or whether the difference actually exists in the population, i.e., is statistically significant. Before proceeding any further let us see what is meant by *statistical significance.*

In short, a difference is statistically significant if it actually exists in the population. Thus, if a certain city contains 50.5 per cent females and 49.5 per cent males, this is a *statistically* significant difference in the sex ratio of that city's population; no question of sampling variation arises at this point because the percentages refer to the entire population, not to a sample. Suppose, now, a sample of 100 people taken at random in the city contains 53 males and 47 females. The question then arises whether this preponderance of males in the sample is statistically significant. In other words, is it very likely that 53 males out of a sample of 100 people could have been selected from a population actually containing an equal or greater proportion of females, or could this difference only have occurred in a preponderantly male population? If the latter is true, then the observed difference is *statistically significant,* thereby leading to the conclusion that the population actually contains more males than females; if the former is true, then the difference is *not statistically significant,* meaning that a sample of 100 people containing 53 per cent males could easily have been drawn from a population actually containing as many or less males than females purely as a result of random sampling variations.

Now, the purpose of statistical significance tests is to set up criteria and methods of approach for appraising the statistical significance of observed differences. The general approach to the problem is to determine a *region of acceptance* about the hypothetical or actual population value to be tested—an interval over which the corresponding values of similar samples taken from the same population may be considered to fluctuate as a result of random sampling influences. In other words, a sample whose representative statistic falls within this interval may be considered to belong to the same population as any other sample whose statistic falls within the same interval, the difference between the sample and population values being attributed to discrepancies caused by chance sampling variations. The area outside the region of acceptance is termed the *region of rejection,* and the samples whose statistic lies within the region of rejection are considered to be "significantly" different from the population under consideration. . . .

In practice, the region of acceptance is computed as a certain multiple of the standard error. Thus, for a large sample (drawn from a more or less

normally distributed population), a 0.95 confidence coefficient is obtainable by computing the region of acceptance as the real or hypothetical population mean plus and minus 1.96 times the standard error of the sample statistic; [5] the region of acceptance with a 0.99 confidence coefficient is computed as the population mean plus and minus 2.58 times the standard error of the sample statistic, etc.

Suppose, for instance, that a radio sample of several hundred families reveals that 10 per cent of these families listen to a particular program, and it is desired to know whether the true population figure might conceivably be as high as 14 per cent, i.e., whether the proportion of all families listening to this program might actually be 14 per cent, the 4 per cent difference being attributable to random sampling fluctuations. Suppose, further, that by applying the appropriate formula the standard error of the estimate comes out to be 1.5 per cent. With a confidence coefficient of 0.95, the region of acceptance around the hypothetical population value of 14 per cent is computed to cover the interval from 12.5 per cent and upward.[6] Since the sample rating of 10 per cent is beyond the lower limit of the region of acceptance, the conclusion is that this difference is too great to have been caused by random sampling elements, and it is very unlikely that the true proportion of families listening to this radio program is as high as 14 per cent.

A different line of reasoning sometimes employed to reach the same result is to consider the difference between the real or hypothetical population value and the relevant [7] limit of the region of acceptance as constituting the maximum size of the difference that might be attributed to sampling fluctuations. If the difference between the two values to be tested is equal to or less than this allowable maximum, it is adjudged to be not significant; otherwise the difference is held to be a valid change. Thus, in the above example any sample yielding a listenership percentage more than 2.5 per cent below the hypothetical value of 14 per cent would be considered to indicate a significant difference in program listenership. The sample cited above does represent such an instance. As will be shown later, the second method is preferable because of its wider applicability.[8]

The theory of significance tests is not restricted to the testing of the importance of the difference between single values, but is also employed to test the significance of the difference between two or more entire distributions, as in determining the significance of regional differences in consumer income purchase patterns—by means of chi-square and variance analysis—as well as for many other purposes. . . .

Standard Errors and Confidence Re-

[5] Alternatively, it may be obtained as the population mean plus *or* minus 1.645 times the standard error of the sample statistic, or in any other number of combinations.

[6] The mechanics of computation of such intervals is illustrated below.

[7] Depending on whether the value of the other sample is above or below that of the first sample.

[8] Especially when the problem involves testing the significance of the difference between two samples.

gions. It was noted previously that the function of the standard error in the process of statistical estimation is to provide an interval within which the sample statistic might have deviated from the true population statistic as a result of random sampling variations—this is the confidence region, the interval that is believed to contain the true population value. By fulfilling this function, the standard-error concept is at the same time serving its purpose in the theory of testing hypotheses, for the interval that forms the confidence region in statistical estimation corresponds to the region of acceptance in testing hypotheses.[9]

Both regions are based on the standard-error concept and delineate intervals where random sampling fluctuations are thought to cause sample statistics to deviate from the true population value. Whereas in estimation this area is believed to contain the true population value, in testing hypotheses this region is taken to be the area within which similar samples from the same population would fall as a result of chance variations in sampling. Thus, in a survey of the Southwest region, it may be found that 20 per cent of the sample purchases brand *X* coffee. By applying the standard-error formulas, the confidence region (the interval within which the true population value is believed to lie) might turn out to be 17 to 23 per cent, with a probability, i.e., confidence coefficient, of 0.95. If one wishes to ascertain whether this brand is definitely more popular in the Southwest than in the Pacific region, where a similar sample reveals the proportion of families purchasing this brand of coffee to be 16 per cent, the region of acceptance is computed as a weighted average of the standard errors of the two samples. The resultant interval is then taken to indicate the maximum permissible difference that could occur between the two sample averages as a result of random fluctuations.

This, then, is the dual function of the standard-error concept in sampling analysis. It serves to delineate the area of the final estimate and to provide the means of computing the necessary criterion for the determination of the significance of an estimate. . . .

[9] Assuming that the same confidence coefficients are used throughout.

✦✦✦✦✦✦✦✦✦✦✦✦ ESTIMATION AND THE CONSTRUCTION OF CONFIDENCE LIMITS

HARPER W. BOYD AND RALPH L. WESTFALL

After selecting a simple random sample, how does one estimate universe values from the sample data? There are many such values which might be of interest, but the treatment here will be restricted to the

Marketing Research: Text and Cases, *R. D. Irwin, Inc., 1956, 287–297.*

estimation of the two most commonly used, the *arithmetic mean* and a *percentage.*

SAMPLE VALUES AS ESTIMATES OF UNIVERSE VALUES

Assume one is interested in the total sales of Colgate toothpaste in Detroit grocery stores during a given week. For each of the grocery stores the value of this characteristic (sales of Colgate toothpaste) could be measured. The sum of these values would be the total sales of Colgate's in Detroit grocery stores during the given period. The arithmetic mean would be the average sales of Colgate toothpaste per store for the given week, i.e., the total sales divided by the number of stores.

Another parameter of common interest is the *percentage,* or proportion, of items in the universe possessing a particular characteristic. For example, in the toothpaste study one might be interested in the percentage of grocery stores stocking Colgate toothpaste or the percentage of self-service stores stocking Colgate's.

How does one estimate a universe mean from sample data? [1] Intuition suggests that the mean of the sample might be a good estimate of the universe mean. In this, intuition tends to be correct. If all possible simple random samples of a given size are drawn from a universe and the mean computed for each sample, the average of these sample means will be the same as the universe mean. This indicates that the sample mean is an unbiased estimate of the universe mean, i.e., sample means do not tend, on the average, to be higher or lower than the universe mean. This does *not* mean that each sample mean will be exactly equal to the universe mean. Only rarely will this happen; in general, the two will differ.

It may be helpful to illustrate this lack of bias in sample means as estimates of a universe mean. Consider a universe of four items with the values of $A = 1$, $B = 3$, $C = 4$, and $D = 8$. Assume all possible samples of size 3 are selected from this universe. The possible sample means would be:

Sample	*Sample Mean*	*Sample*	*Sample Mean*
A,B,C	$2\frac{2}{3}$	*A,C,D*	$4\frac{1}{3}$
A,B,D	4	*B,C,D*	5

In this case only one of all the possible sample means is equal to the universe mean, but the average of all possible sample means

$$\frac{2\frac{2}{3} + 4 + 4\frac{1}{3} + 5}{4}$$

is equal to the universe mean of 4. The sample mean, then, affords an unbiased estimate of the universe mean, but any one sample mean is unlikely to be exactly the same as the universe mean.

Since the sample mean can be used as an estimate of the universe mean, it offers a method of estimating the universe aggregate or total. If the sample mean is multiplied by the number of items in the universe, the result is an unbiased estimate of the universe total, but it is not likely to

[1] The discussion here is centered on the mean because it is the most widely used measure of central tendency. Other such measures, particularly the median, are sometimes important.

be exactly the same as the actual total.

The percentage of universe items having a certain characteristic is but a special case of an arithmetic mean. Hence, a sample percentage can be used as a direct estimate of a universe percentage. For example, in a sample of 200 drugstore owners, it was determined that 150 preferred to buy pharmaceuticals direct from the manufacturer, i.e., 75 per cent. This percentage is an unbiased estimate of the universe percentage which prefers to buy direct.

INTERVAL ESTIMATION

Different samples from the same universe will give different estimates of the universe value. The estimate obtained from a particular sample will differ from the universe value because of "sampling error"; that is, because the sample selected by chance is not exactly representative of the universe. If the researcher took another random sample from the same universe, the resulting estimate might differ a little, somewhat, or a great deal from the estimate he obtained from his first sample. He is then faced with the problem of determining how precise, or reliable, his sample estimates are.

The investigator would like to determine a range of values within which he can be fairly sure that the true value lies. That is, he would like to be able to construct an *interval estimate*. Fortunately, the theory of simple random sampling (and of probability sampling in general) provides methods for establishing such a range or interval estimate, thereby permitting evaluation of the reliability of sample estimates. Thus, with simple random sample data it is possible to measure the sampling error associated with such estimates as the mean or a percentage, and to set bounds within which the universe value being estimated will likely fall. This is the great advantage of probability samples over nonprobability samples.

THE SAMPLING DISTRIBUTION CONCEPT

It has been emphasized that different samples from the same universe will lead to different estimates of the universe mean or a universe percentage. For example, assume one chooses all possible simple random samples of two each from a college student universe of six students. Each student has monthly income as follows: $A = \$20$, $B = \$80$, $C = \$100$, $D = \$100$, $E = \$100$, and $F = \$200$. The universe mean is \$100. Sample means ranging from \$50 (Sample AB) to \$150 (Samples CF, DF, and EF) would be obtained. The 15 possible sample means would occur as follows:

Sample Mean	*Relative Number of Occurrences*	*Sample Mean*	*Relative Number of Occurrences*
\$ 50	1 out of 15	\$110	1 out of 15
\$ 60	3 out of 15	\$140	1 out of 15
\$ 90	3 out of 15	\$150	3 out of 15
\$100	3 out of 15		

Such a listing of the possible random sample means together with their relative frequencies of occurrence is called the *random sampling distribution of the mean* for samples of two each from the given universe. Any such distribution of sample values under random sampling is called a sampling distribution.

Given a sampling distribution, one can predict the average behavior of the sample estimate under study. In this case, for example, if repeated simple random samples of two each were drawn from the universe of six students and a sample mean computed each time, on the average 60 per cent (9 out of 15) of these sample means would lie between $60 and $100, inclusive. Again on the average, about 73 per cent (11 out of 15) of the sample means would be within $40 of the universe mean ($100), and so on.

The fact that a knowledge of the sampling distribution makes it possible to predict the sampling behavior of the mean or a percentage is of fundamental importance in statistical inference. If one knew only that a particular sample estimate would vary under repeated sampling and had no information as to *how* it would vary, then it would be impossible to devise a measure of the sampling error associated with that estimate. Since the sampling distribution of an estimate describes how that estimate will vary with repeated sampling, it provides a basis for determining the reliability of the sample estimate.

SAMPLING DISTRIBUTION OF THE MEAN IN LARGE SAMPLES

In practical sampling work the composition of the universe from which the sample is drawn is seldom known. Therefore, the sampling distribution of an estimate based on observations from the universe will also be unknown. On the face of it this would seem to rule out the possibility of constructing interval estimates of the type discussed above.

Although this fact does make it impossible to make estimates whose sampling properties are known *exactly,* it does not preclude the making of estimates, the sampling behavior of which is known approximately. Mathematicians have derived a theorem called the Central Limit Theorem, which makes it possible to construct interval estimates whose properties are known sufficiently well for most practical purposes. In rather crude terms, the Central Limit Theorem states: The sampling distribution of the mean, for a large sample, will be approximately a *normal distribution.*

This large sample distribution of sample means will be distributed symmetrically around the true universe mean as shown in Figure 1. The sample means tend to cluster around the universe mean. Small deviations from the universe mean are more frequent than large ones. Sample means that deviate very widely from the universe mean are rare. Positive and negative deviations of equal magnitude occur with equal frequency.

If a number of samples of a large size is drawn at random from the same universe, the means of the samples will tend to form a normal curve around the universe mean. The larger the individual samples, the more closely will the sample means cluster around the universe mean. That is, the larger the

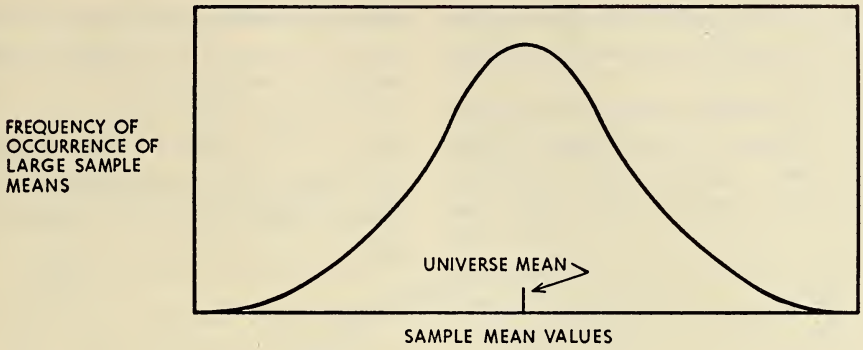


FIGURE 1

DISTRIBUTION OF LARGE SAMPLE MEANS AROUND UNIVERSE MEAN

sample, the greater the reliability of the sample mean. This accords with common sense since it would be expected that a large sample would be more similar to the universe than would a small sample.

larger samples has fewer values that deviate widely from the universe mean. The larger the samples, the more closely the sample means will cluster around the universe mean. However, no matter how large the sample, unless it ap-

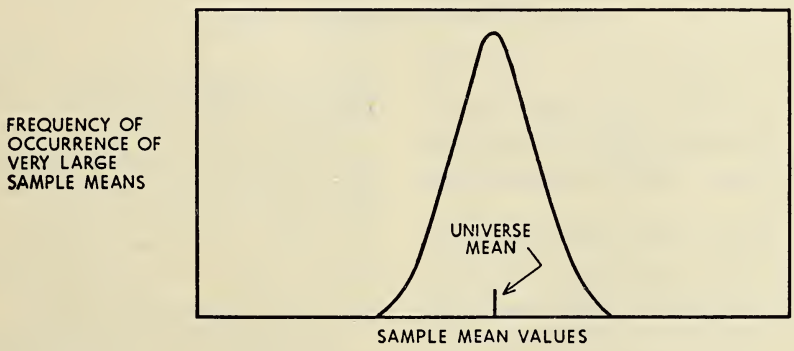


FIGURE 2

DISTRIBUTION OF SAMPLE MEANS AROUND UNIVERSE MEAN FOR VERY LARGE SAMPLES

Figure 2 illustrates the closer clustering of sample means around the universe mean when very large samples are used. Compare this distribution with that shown in Figure 1. Note that both are distributed symmetrically, but the distribution of sample means for

proaches the size of the universe, there will be some sample means which are different from the universe mean.

If an entire universe is distributed normally around its mean, the proportion of the universe included between two limits is determined by the devia-

tions of those limits from the universe mean, measured in terms of standard deviations.[2]

In a normally distributed universe, about 68 per cent of the items will be within one standard deviation of the mean, about 95 per cent within two

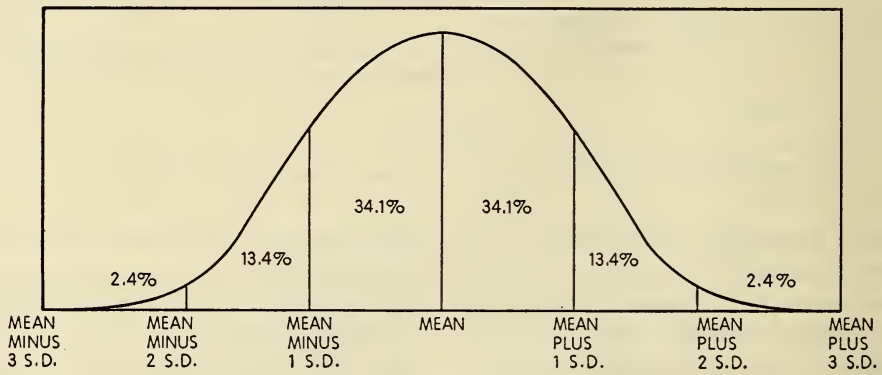


FIGURE 3

AREA UNDER THE NORMAL CURVE

standard deviations of the mean, and virtually all (99.7 per cent) within three standard deviations of the mean. These facts are exhibited geometrically in Figure 3.

Since sample means will be approximately normally distributed about the universe mean, it is possible to make analogous statements about the deviation of sample means from the universe mean. When applied to the sampling distribution of the mean, however, the term *standard error of the mean* is used instead of standard deviation. The following statements apply.

a) About 68 per cent of sample means will fall within one standard error on either side of the universe mean.

b) About 95 per cent of sample means will fall within two standard errors on either side of the universe mean.[3]

c) Practically all sample means will be located within three standard errors on either side of the universe mean.

This leaves the problem of determining the actual size of the standard error

[2] The standard deviation of a universe is a measure of the dispersion of the items in the universe around their mean. The standard deviation of a finite universe may be determined by the formula

$$\sigma = \sqrt{\frac{\Sigma x^2}{N-1}}$$

Where

σ = standard deviation

x = deviation of an item from the universe mean

N = number of items in the universe

where N is the universe size and n is the sample size.

[3] Technically, 95 per cent of the sample means will fall within 1.96 standard errors on either side of the universe mean. The 1.96 is rounded to 2 to simplify the illustrations used in this section and to conform to common usage.

of the mean. This is obtained by the following formula:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Where

$\sigma_{\bar{x}}$ = standard error of the mean
σ = standard deviation of the universe
n = number of observations in the sample

This formula applies if less than 5 per cent of the universe is included in the sample.[4]

Nothing has been said as to how large the sample must be before the Central Limit Theorem may be applied. The question can be answered only by a knowledge of the universe being sampled. For fairly symmetrical universes, the sample can be as small as 10 and the approximation will hold very well. On the other hand, for very skewed universes (which are not uncommon in marketing), the approximation may be a relatively poor one, even for a sample as large as several hundred. For most applications it is probably satisfactory to assume that means based on samples of 30 will be approximately normally distributed.

CONSTRUCTION AND INTERPRETATION OF CONFIDENCE INTERVALS

How does the researcher construct an *interval estimate* of the universe mean? This procedure will be illustrated by the construction of what is called a 95 per cent *confidence interval* estimate of the mean. Consider the following hypothetical situation. The mean of a certain universe (M) is unknown, but the standard deviation of that universe (σ) is known. A sample mean which is based on a sample of $n = 100$ observations is available.

Our knowledge of the sampling distribution of the mean tells us that the interval

$$M \pm 2\frac{\sigma}{\sqrt{n}} = M \pm 2\frac{\sigma}{\sqrt{100}} = M \pm 2\frac{\sigma}{10}$$

will include about 95 per cent of all possible sample means of samples for which $n = 100$. That is, 95 times out of a 100, we will be right if, *before* the sample of 100 is chosen, we assert that the sample mean we will obtain will lie

between $M - 2\frac{\sigma}{10}$ and $M + 2\frac{\sigma}{10}$.

After the sample is known, we *assume* that the sample mean actually obtained is one of those which does lie in this interval. If this is true (and by definition it will be true in 95 per cent of repeated samplings), then the sample mean will be located between

$M - 2\frac{\sigma}{10}$ and $M + 2\frac{\sigma}{10}$. Denoting the sample mean that was actually obtained by $\bar{x}$, the statement: "M lies between

$\bar{x} - 2\frac{\sigma}{10}$ and $\bar{x} + 2\frac{\sigma}{10}$" will be correct 95 per cent of the time, if we draw a large number of samples.

In a practical situation, a particular

[4] If more than 5 per cent of the universe is included in the sample, then the standard error becomes

$$\sqrt{\frac{(N-n)}{N}} \cdot \frac{\sigma}{\sqrt{n}}$$

where N is the universe size and n is the sample size.

sample mean will have been obtained so that we can solve numerically the two equations:

$$M = \bar{x} - 2\frac{\sigma}{10}$$

$$M = \bar{x} + 2\frac{\sigma}{10}$$

This will give us a range of values within which the population mean M may, with 95 per cent confidence, be presumed to lie. More precisely, if a large number of random samples of size 100 is drawn from this universe and each time the statement is made that M lies between $\bar{x} - 2\frac{\sigma}{10}$ and $\bar{x} + 2\frac{\sigma}{10}$ (where $\bar{x}$ is computed afresh for each sample), then 95 per cent of these statements will be correct. This is the meaning of the expression "95 per cent confidence interval."

By the same reasoning, a confidence interval which is "almost certain" to cover the universe mean may be computed using the limits:

$$M = \bar{x} - 3\frac{\sigma}{\sqrt{n}}$$

$$M = \bar{x} + 3\frac{\sigma}{\sqrt{n}}$$

(Three standard errors are used because almost all of the observations in a normal distribution are located within three standard deviations of the universe mean.)

It must be emphasized that this interpretation of "confidence" in a particular interval estimate is defined in terms of what will happen if a large number of samples is drawn. Any particular confidence interval either will or will not cover the universe mean. After the sample has been drawn it is a matter of fact whether or not the particular interval estimate does cover the universe mean. What is guaranteed is that, for example, 95 per cent of such statements about the location of the universe mean will be correct if a 95 per cent confidence interval is used.

Up to this point it has been assumed that the standard deviation of the universe is a known quantity. In practical . . . problems this value will almost never be known. Therefore, an estimate must be substituted in the formula if the standard error is to be estimated. The standard deviation of the sample is used for this purpose resulting in the estimated standard error

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Where

$s_{\bar{x}}$ = estimated standard error of the mean
s = standard deviation of the sample
n = number of observations in the sample

This estimate tends to distort the accuracy of the estimate of the confidence limits, but, for practical purposes it will be satisfactory if the sample is large.

An illustration may be of value in illustrating the concepts expressed relative to confidence intervals.

Example 1. A random sample of 400 housewives provided the following information about the amount of money

spent during a six-month period on food items:

Sample mean $(\bar{x}) = \$400$
Sample standard deviation $(s) = \$80$

The problem is to estimate the mean of the universe with a 95 per cent confidence interval and to interpret the result. The algebraic expression for this confidence interval is

$$\bar{x} \pm 2\left(\frac{\$80}{\sqrt{400}}\right) = 400 \pm 2\left(\frac{\$80}{20}\right) = 400 \pm 2(\$4) = 400 \pm \$8.$$

We are 95 per cent confident that the universe mean will be between \$392 and \$408. This finding is interpreted in the following way: If a very large number of samples of 400 housewives each was selected at random from this universe and for each such sample the confidence interval $\bar{x} \pm 2\frac{s}{\sqrt{n}}$ was computed, then about 95 per cent of the intervals so computed would include the universe mean.

It is to be emphasized that the interpretation did not say the particular interval, \$392 — \$408 (as calculated from a single sample), will bracket the universe mean. Rather the interpretation said that, if a large number of random samples were selected and the interval computed each time, then about 95 per cent of the intervals would cover the universe mean.

CONFIDENCE LIMITS FOR PERCENTAGES

As in the case of the universe mean, the researcher may also wish to construct confidence limits for population percentages. Fortunately, the theory is identical to that used to construct confidence limits for the universe mean since a percentage is but a special case of the mean. It follows that the sampling distribution of a percentage is, for large samples, approximately normally distributed. The standard error of a percentage, σ_p, from a simple random sample is estimated by the formula $\sqrt{\frac{pq}{n}}$, where p is the percentage of items in the sample possessing a given characteristic, q is the percentage of items not possessing the characteristic, and n is the sample size. Assume that a simple random sample of 100 families shows 40 have a TV set and 60 do not—i.e., 40 per cent have TV. The estimated standard error, s_p, would be computed as follows:

$$s_p = \sqrt{\frac{.40 \times .60}{100}} = \sqrt{\frac{.24}{100}} = \sqrt{.0024} = 4.9\%$$

The 95 per cent confidence interval would be 2(4.9 per cent) = 9.8 per cent. Thus, we would be 95 per cent confident that the universe percentage of families having a TV set was within 40 per cent ± 9.8 per cent or between 30.2 per cent and 49.8 per cent.

One caution is in order with regard to this method of constructing a confidence interval for percentages. For values of p less than 30 per cent or more than 70 per cent, a sample of more than 100 is needed if the normal approximation to the sampling dis-

tribution of a percentage is to be a good one. If, for example, one were sampling an infrequent attribute, one which only 2 per cent of the universe had, he could not realistically assume that the sample percentage would be distributed normally under repeated random sampling unless a large sample (more than 100) was used.

✦✦✦✦✦✦✦✦✦✦✦✦✦ TESTS OF SIGNIFICANCE

HARPER W. BOYD AND RALPH L. WESTFALL

Such tests are used to determine whether the difference between two or more statistics is significant. The statistics may derive from the same survey, they may come from different surveys, or one may be a statistic while the other is a hypothetical value. Tests of significance are designed to determine what part of the difference may be attributed to chance or random sampling variations. Thus, a sample mean which varied from another sample mean by one standard error of the difference would not be thought of as differing significantly, because of the relatively high probability that this deviation occurred due to sampling variation. But if the difference between the two values were 3 standard errors, then there would be only a very slight chance (3 out of 1,000) that the difference was not significant. Just what level of significance the analyst uses is an individual matter. Thus, some use the 95 per cent level of confidence (2 standard errors) which means that only about 5 times out of 100 could the difference be due to chance. The confidence level selected depends essentially on the importance of the decision to be made.

The basic formula used in tests of significance is

$$\frac{\text{Sample Value A minus Sample Value B}}{\text{Standard Error of the Difference between A and B}}$$

Sample values A and B are comparable values (e.g., percentages) from two independent samples, two subsamples, or from the same sample if the two values are independent of each other, i.e., the size of one does not determine the size of the other. The measure derived from this formula is checked with an appropriate probability distribution table to determine the probability of the difference being significant. There are several different formulas which may be used in particular cases. The correct one to use depends upon what differences are being tested, the type of sample estimate involved, and the size of the sample. Thus, different formulas are used for determining whether significant differences exist between

Marketing Research: Text and Cases, *R. D. Irwin, Inc., 1956, 438–440.*

two sample means and between two sample percentages.[1]

A CASE ILLUSTRATION

A 1952 study was made in a Midwestern city among supermarkets to determine the per cent of stores stocking various brands of frozen orange juice. The results were:

Brand	*Per Cent of Stores Stocking* *
A	48
B	21
C	55
D	15
Other	62

* Percentages based on reports from 100 stores.

These results were obtained using a random sample. The question to be answered is whether the per cent of stores stocking brand C (55%) is *significantly* greater than the per cent stocking brand A (48%). The standard error of the difference formula for the difference between two percentages is

$$s_{difference} = \sqrt{s_a^2 + s_b^2}$$

Where:

$s_{difference}$ = standard error of difference
s_a = standard error of percentage A
s_b = standard error of percentage B

[1] For a thorough discussion of these and other alternative significant difference formulas, see Robert Ferber, *Statistical Techniques in Marketing Research* (New York: McGraw-Hill Book Co., Inc., 1949), pp. 112–28.

The standard error of a percentage is determined by the formula:

$$s = \sqrt{\frac{pq}{n}}$$

Where:

s = standard error
p = percentage with characteristic under study
q = 100 per cent minus p
n = size of sample

Substituting in the previous formula:

$$s_{difference} = \sqrt{\left(\sqrt{\frac{p_a q_a}{n_a}}\right)^2 + \left(\sqrt{\frac{p_b q_b}{n_b}}\right)^2}$$

$$= \sqrt{\frac{p_a q_a}{n_a} + \frac{p_b q_b}{n_b}}$$

$$= \sqrt{\frac{.55 \times .45}{100} + \frac{.48 \times .52}{100}}$$

$$= \sqrt{\frac{.2475}{100} + \frac{.2496}{100}} = \sqrt{\frac{.4971}{100}}$$

$$= \sqrt{.004971} = .0705 = 7.05\%$$

This figure is then divided into the difference between the two percentages:

$$\left(\frac{7\%}{7.05\%}\right) = 1 \text{ standard error.}$$

The probability of such a difference occurring by chance even though no real difference existed is 32 out of 100, and thus it is concluded that the difference between brand C and brand A is not very significant, i.e., the observed difference could have occurred as the result of chance fairly easily.

In 1953 a similar study using a sample of 200 instead of 100 was made in the same city with the following results.

Brand	*Per Cent of Stores Stocking* *
A	47
B	26
C	63
D	8
Other	58

* Percentages based on reports from 200 stores.

The question here is whether the change in brand C from 55 per cent of stores stocking to 63 per cent is significant. The standard error of the difference formula for percentages in this case is:

$$\sqrt{\frac{.55 \times .45}{100} + \frac{.63 \times .37}{200}}$$

$$= \sqrt{\frac{.2475 + .2331}{100 \quad 200}}$$

$$= \sqrt{.002475 + .00117} = \sqrt{.00365}$$

$= .0604$ or 6.04%.

This figure is divided into the observed difference $\left(\frac{8\%}{6\%}\right)$ and the result is 1.33 standard errors. This difference could occur by chance 18 times out of 100 even though there was no actual difference in the percentage of stores stocking the brand in the two years. This means the difference is more significant than the previous one but still would not be considered highly significant.

The same procedure, but using a different formula, would be followed if the sample estimate to be checked was the arithmetic mean instead of a percentage. It is important to note that all tests of significance are based on the fact that a probability sample was used.

✦✦✦✦✦✦✦✦✦✦✦✦ THE TESTING OF HYPOTHESES

E. BRIGHT WILSON

Experiments are often carried out in order to test hypotheses. For example, there may be reasons which suggest that antihistamine drugs will cure colds. On the other hand, it is well known that individuals differ in their susceptibility to colds, that colds are hard to define, and that psychological influences are important. Consequently, it is expected that the results of any test will be somewhat variable at best. An apparently successful test can easily represent a mere chance combination of circumstances.

The hypothesis that the results obtained were due to the chance effects of uncontrolled variables is an exam-

ple of what is called a *null hypothesis.* In general, the null hypothesis is always in a form which permits the calculation of the probability of each possible result.

The hypothesis that the drug is effective to a certain definite extent is an example of what is called an *alternative hypothesis.* The analysis of the results of an experiment may therefore be a question of deciding whether the null hypothesis or an alternative hypothesis provides the more plausible explanation of the observations actually obtained.

In more general situations there may be several alternative hypotheses. For example, the investigator may feel that there are reasons why the administration of the antihistamine drug may interfere with the patient's natural recovery. He would therefore wish to examine the results of any test to see whether or not they supported this view. It would be natural under most circumstances to consider both possibilities, i.e., that the drug might be beneficial and that it might be harmful. Another somewhat unusual alternative hypothesis would be that someone had tampered with the results, consciously or unconsciously, so as to suppress any differences between subjects and controls.

More complicated alternative hypotheses are possible. For example, it might be suspected that women were more affected by the drug than men, or that children reacted more than adults, or thin people more than fat people. It does not require much extension of these examples to see that an alternative hypothesis can generally be invented—after the experiment—which will provide a perfect fit for any possible result. It is therefore logically unsound to use the results of an experiment as *proof* of an alternative hypothesis which was conceived after the results were at hand. However, such situations may legitimately be used to suggest an alternative hypothesis to be tested by the next experiment, and this is a very important method for producing new hypotheses.

ERRORS OF THE FIRST KIND

Suppose that, before making an experiment on the effect of a drug, two alternative hypotheses are proposed: that it will have a noticeable effect, either consistently beneficial or consistently harmful. As a concrete example, let the experiment consist in administering the drug to five subjects, five others serving as controls. For simplicity suppose that the subjects and controls have been matched in pairs. It is assumed that all . . . precautions . . . have been taken, such as randomization, use of placebos, and blindfoldedness. Now consider a hypothetical case in which four of the subjects recover more rapidly than their controls, while the reverse is true for the fifth subject. In other words, the drug appears to have been successful in four out of five cases. How should this result be interpreted?

Probably most persons would regard four positive results out of five tries as rather convincing evidence against the null hypothesis and thus as strong support for the alternative hypothesis that the drug was effective. It will be shown that under most circumstances this is an erroneous conclusion.

Is it reasonable to suppose that such a result could have occurred by chance? Suppose the null hypothesis to be true so that subjects reacted differently from controls because they were different anyway, and not because of the drug. Since the choice of subject from each matched pair was made by tossing a coin, the mathematics of coin tossing can be applied to the problem. This is one of the several benefits of randomization. Suppose that, in the first four pairs, the subject came out best. If this were sheer luck, the result of the throws of the coin in the randomizing operation, it would have a probability of $(1/2)^5$, or 1/32. One might perhaps regard this as too small to be likely to have happened by chance. But here the great apparent paradox of probability enters. The probability of every detailed event, taken by itself, is always small; yet such events do happen by chance.

What is the probability that the top dollar bill in a wallet has the serial number 56078727? It is certainly exceedingly small. Examination shows that the number is actually 84584017. What was the probability, before the event, of this number appearing? Just as minute as any other, and yet it did occur. Therefore, it certainly will not do to say that an individual result cannot have a chance origin just because its probability is low.

To return to the experiment, there are thirty-two possible detailed results, shown in Table 1. On the basis of the null hypothesis and the randomization process, each of these is equally likely. One of the thirty-two was necessarily obtained in the experiment. Did it arise as a result of chance or as a result of the drug? There is no certain way of deciding. If the investigator accepts the result as being due to the drug, he is taking a risk of drawing a wrong conclusion. Anyone who takes risks a large number of times must expect some failures. Therefore it is inevitable that if scientists always make such decisions on data such as this, in the long run a certain fraction of their conclusions will be erroneous. How big is that fraction?

The risk α of drawing an erroneous conclusion of this kind has been called the risk of an *error of the first kind.* It is the danger of falsely rejecting the null hypothesis. In the example under discussion its magnitude is easily calculated. If the result actually obtained —the first 4 plus, the 5th minus—was accepted as proof of drug action, then certainly any of the other 4 out of 5 situations would likewise have been accepted. Also, the stronger result, 5 out of 5, would have been similarly interpreted. This is a total of 6 cases. But since it was considered possible that the drug might retard recovery, if 5 negative results, or 4 out of 5 negative results, had occurred, it would have been concluded that the drug was active but deleterious. Therefore, out of 32 possible cases, all equally likely on a chance basis, a total of 12 would have been interpreted in favor of drug action. Consequently, even if a nurse has mistakenly given placebos instead of the drug to the subjects as well as to the controls, in the long run 12/32 (or more than one 1/3) of the time such an experiment would yield a result at least as convincing as 4 out of 5. This risk of an error of the first kind is clearly too

high to be acceptable in most circumstances.

TABLE 1

THE THIRTY-TWO POSSIBLE RESULTS OF AN EXPERIMENT WITH FIVE PAIRS

(*A + sign means that subject exceeded control*)

Pair: 1	2	3	4	5	*No.* +
+	+	+	+	+	5
+	+	+	+	−	4
+	+	+	−	+	4
+	+	−	+	+	4
+	−	+	+	+	4
−	+	+	+	+	4
+	+	+	−	−	3
+	+	−	+	−	3
		8 others with 3 +			
+	+	−	−	−	2
		9 others with 2 +			
+	−	−	−	−	1
		4 others with 1 +			
−	−	−	−	−	0

The risk level which would be unacceptable as the basis for a firm decision or for any action which was dangerous or expensive may be entirely adequate for suggesting the continuation of an experiment. In fact, it is probably true that the first evidence leading to important scientific discoveries is often very slender and could be accepted as proof only at very great risk. Therefore it is unwise to set α too low in the preliminary stages of an investigation. Later on, however, when it is a question of asserting that something has been proved, more rigorous criteria are usually necessary.

In the above case the risk α could have been reduced to 2/32, or 1 in 16, by accepting only 5 out of 5 (either way) as evidence of drug action. Better still, more experiments could be run. If 20 pairs were tested and any result at least as conclusive as 16 out of 20, either way, accepted as proof of drug action, the risk α would be very much lower, even though the proportion is the same: 4 out of 5. There would here be 2^{20} equally likely results (on the basis of the null hypothesis), each of which therefore has the very small probability of 2^{-20}. There are 12,392 possible results showing 16, 17, 18, 19, or 20 positive out of 20, or 16, 17, 18, 19, or 20 negative out of 20. Therefore, if there were no drug action but only chance variations between subjects and controls, the probability of getting one of these critical results is $12{,}392 \times 2^{-20} \cong 0.012$, a very considerable reduction in risk.

The risk level α is often called the *level of significance.* Experiments are said to be *statistically significant* if their result, plus all the other possible results at least as convincing as the one obtained, under all the alternative hypotheses specified ahead of time, have a total probability of α or less on the basis of the null hypothesis. Naturally the value of α must be specified.

ERRORS OF THE SECOND KIND

An experiment is not properly planned merely because the risk α is kept low. If this were so, no experiments would ever be necessary. Hypotheses would be tested by spinning a roulette wheel. A fraction α of the rim would be marked "Reject Hypothesis." Since this would come up only a fraction α of the time, correct null hypotheses

would be wrongly rejected at most a fraction α of the time in the long run. It is obvious that something more is needed. The trouble with the above scheme is that it also has a small chance α of . . . [accepting] the null hypothesis when it ought to be rejected because the alternative hypothesis is the correct one. No reputable scientist wishes to announce that he has made a discovery and then later have it shown that his conclusion was erroneous. He therefore will keep the risk α low. But it is also important not to miss positive effects that are real. The risk of dismissing such real effects is denoted by β and is called the chance of an error of the second kind. Here one mistakenly rejects the alternative hypothesis.

The two quantities α and β are to a certain extent antagonistic. By requiring more and more stringent evidence in an experiment of given size (for example, requiring 5 out of 5 instead of 4 out of 5), α can be reduced, but this automatically increases β. However, by enlarging the experiment, both risks can be decreased. Furthermore, β can often be decreased by improved experimental techniques. In the example of the cold cure, these could include more careful matching of subjects and controls, perhaps based on records of their susceptibilities in the past. The technique of examining the patients might also be improved so as to detect smaller differences.

POWER OF A TEST

In a given situation, the choice of the risk α will be governed by the penalty exacted for making this kind of error and by other considerations. As an illustration, let it be taken as 1.2 per cent. If the size of the experiment is fixed, this fixes the *number* of results which can be used as evidence for the alternative hypotheses. With 20 pairs, it will be 12,392. As far as the risk α of being deceived by chance effects is concerned, it will make no difference at all which 12,392 possible results are set aside as *critical,* i.e., to be used as evidence for a real effect (or more generally as evidence against the chance origin under the null hypothesis).

The choice of the critical set will, however, make a great difference in the risk β of missing a real effect. Since there will be some penalty attached to failure to find regularities which are really present, β should be minimized by making the best selection of the critical set. In this connection, statisticians have defined a quantity called the *power* of a test:

$$\text{Power} = 1 - \beta \qquad (1)$$

If a test of a particular alternative hypothesis is based on a critical set of results which makes β a minimum (for fixed α and fixed size of experiment), it will be called a *most powerful test.* The critical set should consist of those results for which the ratio

$$\frac{\text{Prob. on basis of alternative hypothesis}}{\text{Prob. on basis of null hypothesis}}$$

is the greatest. In the example, every result (of the 2^{20} possible) is equally likely on the null hypothesis, but 20 out of 20 positive is more likely than any particular other case if the drug is really curative. This is followed in order by 19 out of 20, 18 out of 20, etc.

When there is more than one alternative hypothesis, there may not be any one critical set of results which leads to a test which is most powerful with respect to all of the alternatives. When there is a critical set which is most powerful for all the alternative hypotheses specified, it is called a *uniformly most powerful set* for that set of hypotheses.

There is another approach to the testing of hypotheses which is based on rather different philosophical ideas. . . . Fortunately, the two methods seldom disagree in practice.

THE PLACE OF STATISTICS

As shown in the above discussion, statistics provides a basis for judgment by enabling the risks of various types of errors to be estimated, but it does not replace judgment. The investigator himself has to decide the levels of risk he will accept. He cannot evade his responsibility for these decisions by hiding behind some conventional level of significance, such as 1 in 20. Each problem is different and has a different background of prior knowledge. Only the experienced investigator knows what are the reasonable alternative hypotheses, what their status is on the basis of prior knowledge, what is the cost of errors of either kind, what the expense of further experiments would be, and whether or not these experiments would be sufficiently free of systematic error so that they would actually increase the precision. The necessity for answering these questions should show the fallacy of expecting a statistician, unfamiliar with the background, to analyze tables of data presented to him as an abstract problem. The statistician can give invaluable advice regarding the planning of an experiment and the methods suitable for the analysis of the data, but he is ordinarily not equipped with the necessary background information to answer alone the questions listed above though he is well aware of their importance.

Furthermore, fifty pages of higher mathematics will not salvage an experiment with a hidden bias. A remarkable example of such a situation occurred in some tests of a proposed medical treatment. The experiment was very carefully designed with randomized controls, placebos, and precautions to keep either the patients or the examining physicians from knowing which were subjects and which controls. For several weeks the treatment showed excellent results, with very satisfying statistical significance. Then one day there was an obvious change, and thereafter only random fluctuations separated the subjects from the controls. The most searching examination failed to find the cause of this shift until finally it was learned that the receptionist on duty during the first weeks had started her vacation on the critical day. Further inquiry showed that she had known which of the patients had received the treatment. She thoroughly believed in it, with the result that her cheery greetings to the lucky subjects, "How much better you look this morning, Mr. Smith!" had so brightened them up before they went in to be examined that the reports of the physicians were seriously biased. . . .

RESULTS WHICH APPEAR "TOO GOOD"

The importance of the choice of alternative hypotheses and of their clear statement in advance of the analysis is illustrated by the problem of results which fit the null hypothesis extremely closely. Suppose that a group of students is assigned an exercise in which they are supposed to toss a coin 1,000 times and record the number of heads. If they report that they obtained exactly 500 heads, their teacher might well suspect that they fabricated the results. Here the null hypothesis is that each throw has a 50-50 chance of yielding a head, regardless of the results of previous throws. But what are the alternative hypotheses? Of course, there is some possibility of securing a biased coin, but this is not very likely. A certain number of extreme results, such as 800 to 1,000 heads or 800 to 1,000 tails, could be set aside as a critical set which would cause rejection of the null hypothesis and therefore acceptance of the alternative hypothesis of bias. An experienced teacher would, however, have still another alternative hypothesis in mind, *viz.*, that the results were invented or altered by the students. To test for this, a critical set consisting of results near 500 heads would be the most reasonable, since an inexperienced student would be likely to set his alleged results near this "most probable" value. Naturally, the total size of the critical set must not be too large since its size sets the risk α of an error of the first kind. The division of the critical set between the extremes (near all heads, all tails) and the center (500 heads) is a matter of judgment, which must include estimates of the character of the students. Actually it would be a disreputable group indeed which would call for a center set consisting of more than the one result 500 heads, because this alone will contribute about .025 to α.

Deliberate deception is very rare in science, but self-deception is all too common. If the null hypothesis is dear to the investigator's heart, there are many ways in which he can influence the results of his experiments so as to come closer to the ideal value. He can do this without fully realizing it. Even in the coin experiment, there is room for bias. If heads have come up "too often," the experimenter may be tempted not to count throws which roll under the sofa (if they are heads). In real scientific work the opportunities for the introduction of bias are much more numerous and subtle. Consequently, it is always reasonable to include the existence of bias as one of the alternative hypotheses, even in analyzing one's own work. The size of the critical region to be set aside for this hypothesis depends on the circumstances.

It cannot be repeated too often that the null hypothesis is rejected, not because a result of low probability is obtained, but because an alternative hypothesis of reasonable prior probability has been considered. The sequence heads, heads, tails, heads, heads, tails, tails, heads, tails, tails is quite improbable (1 in 1,024), but it happens to be the first one obtained in a trial. The assertion that the coin was good and the

throw was fair is opposed by no reasonable alternative hypothesis which would include this particular sequence in its critical set (unless the reader wishes to charge the author with fabricating a "too good" example containing 5 heads in 10 throws! However, the risk α would be 24 per cent for this since all 252 of the 5 out of 10 results would have to be included in the critical region). . . .

THE ESTIMATION OF PARAMETERS

The methods of testing hypotheses can answer the question, "Does this cause produce a real effect?" but what is often desired is an answer to the query, "How great an effect does this cause produce?" It is the purpose of methods of estimation to answer this question. The answer can be in either of two forms. An interval, or range of values, may be given which is believed to cover the true value. Alternatively, a single number may be supplied which is alleged to be a good estimate of the true value.

The form of estimate which leads to an interval is illustrated by the method of *confidence intervals*. . . . It will be remembered that a confidence interval is a range of values determined by a procedure which ensures that in some high proportion (say 95 per cent) of similar applications the interval will include the true value. Confidence intervals cannot be calculated for all types of problems but are available for a number of important cases. Thus, the derivation of confidence intervals for the proportion of marked items in a class (the "binomial" problem) is worked out in detail . . . [on the following pages.] . . .

EXPERIMENT AS A SAMPLING PROCESS

In applying statistical arguments to the analysis of experimental data, it is customary to treat the problem as a branch of the theory of *sampling*. The experimental results which are actually obtained are considered to be a sample from the class (or *population*) made up of all the possible results. For example, suppose that the velocity of light is being measured. Then the finite number of values which are obtained are regarded as a sample from a hypothetical population made up of the infinite number of observations which might have been taken. The process of making the measurement or carrying out the observation is thus thought of as equivalent to drawing chips from a bowl.

It is usually assumed that the sample is a random one, but this is very often a mistaken assumption which leads to error. If randomness has been deliberately and carefully introduced, then it should be legitimate to assume randomness in the result, but in many experiments and measurements this is not the case. Wherever possible, randomness should be introduced . . . , but in addition the results should be tested for possible nonrandom influences. . . .

This process of sampling from a hypothetical population is just as susceptible to various types of error as is sampling from real populations. Thus

the sample may be biased in a variety of ways—for example, by the indiscriminate rejection of observations. In addition, sampling variations are inevitable. In the sections which follow, the effect of these sampling variations will be explored in detail for a few important cases.

SAMPLING FOR ATTRIBUTES—THE BINOMIAL DISTRIBUTION

As a concrete example of the modern mathematical treatment of sampling, consider a population of which a proportion p possesses some attribute, say redness, so that the fraction $1 - p$ lacks this attribute. If one individual is drawn at random from the population, it has a probability p of being red and a probability $1 - p$ of not being red. If a second individual is drawn after the first one has been replaced, the second will likewise have probability p of being red. There are three possible results for the sample of two: both red (probability p^2), one red [probability $2p(1-p)$], neither red [probability $(1-p)^2$]. The coefficient 2 in the middle result comes from the two ways in which this event could occur: the red individual being drawn first or being drawn second. . . .

This argument can be extended to samples of n individuals drawn at random, with replacement. The probability of r red things in a sample of n is

$$P_n^r = \frac{n!}{[r!(n-r)!]} p^r (1-p)^{n-r} \qquad (1)$$

where the factorial coefficient is the number of different ways of ordering the r factors p and the $n - r$ factors $(1-p)$. The basic assumption on which this formula is founded needs emphasis. It is that each individual drawn has a probability p of being red, regardless of what the result of the previous drawings has been. This con-

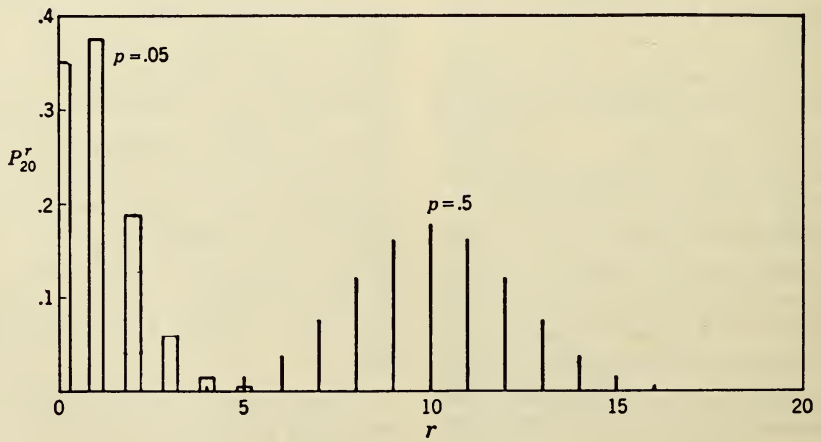


FIGURE 1

The binomial distribution with $n = 20$, $p = .05$, and $p = .5$

dition will be achieved if the sample is a truly random one from a population which does not change during the drawing and if after each draw the item is replaced so that it has an equal chance with every other individual of being drawn again the next time. With small samples from large populations, replacement is not important.

Figure 1 shows some values of P_n^r for two values of p, with $n = 20$. This shows some of the general properties of this distribution, which is called the *binomial distribution*, the properties of which have been elaborately studied and discussed. Table 2 gives values for a special case.

TABLE 2

BINOMIAL DISTRIBUTION FOR SAMPLES OF 20 FROM A POPULATION WITH $p = .05$

r	P^r_{20} [1]	$\sum_{r'=0}^{r} P^{r'}_{20}$ [2]
0	.3584	.3584
1	.3774	.7358
2	.1887	.9245
3	.0595	.9840
4	.0133	.9973
5	.0023	.9996
6	.0003	.9999
7	.0000	.9999

[1] Probability of the number r of red individuals in a sample of 20 (ed.).

[2] This is the cumulative probability (ed.).

PREDICTION OF SAMPLE DISTRIBUTION

This distribution has several applications. In the first place suppose that the proportion p of a certain attribute in a given population is known and it is desired to know the distribution of this attribute in repeated samples of a given size. This is given directly by P_n^r of Eq. (1), which is the probability and therefore the long-run proportion of r red things in samples of n. As an illustration, suppose that a school board proposes to put a few left-handed writing-arm chairs in its classrooms. Very roughly, five per cent of children are left-handed. In classrooms of 20, one would thus expect on the average 1 left-handed child. However, from Table 2, summation of P_{20} from $r = 2$ to 20 shows that 26 per cent of the samples of 20 (if they are really random) will have more than the average number. The school board will thus have to decide what chance it is willing to take of having insufficient chairs because it can be *certain* of having enough only by providing 20. If they agree to take a 1 in 50 chance of having too few, then Table 1 shows that 3 chairs would be sufficient, since the accumulated probabilities of 20, 19, 18, . . . , 4 add to 1.6 per cent.

Now suppose there are 10 such rooms in the whole school, thus calling for 30 chairs on the above basis. A statistically minded board member might point out that if the school is treated as a whole with $n = 200$ pupils, calculations with Eq. (1) [or more easily with Eq. (6)] show that if 17 left-handed chairs are provided there is only a 1 in 50 chance of having too few, a saving of 13 chairs. Therefore it will save money if the chairs can be shifted from room to room. This illustrates that, *on a percentage basis*, fluctuations about the expected value decrease with an increase in sample size, although the actual values of the fluctuations of course increase.

This example also illustrates that some assumptions must be made in applying statistical formulas to actual situations. These assumptions can often not be justified in a rigorous way but rest on judgment. The assumption here is that each classroom has a random sample, as far as left-handedness is concerned, from a population with 5 per cent left-handed children. Clearly it is not random in the general sense; it has been selected on an age and a geographical basis. It is very probable that closer investigation would show strong divergences from the assumption used, but Eq. (1) might still be a good practical basis for the school board's action.

TESTING HYPOTHESES

The next application of the binomial distribution will be to the testing of hypotheses. Suppose it is wished to test the hypothesis that a given schoolroom contains a sample that is random as far as left-handedness is concerned and drawn from a population containing 5 per cent left-handed children. If there are actually 20 out of 20, should the hypothesis be rejected? Surely in this case the hypothesis would be rejected, since 20 out of 20 could arise only once in 10^{26} times on a chance basis if the hypothesis were correct. But suppose the whole city has just 5,000 left-handed pupils in a total of 100,000 pupils, just the expected number. Calculation with Eq. (1) [or with the simpler method of Eq. (6)] shows that the probability of this result is only 1 in 173. Thus the perplexing paradox arises that the expected number (which is also in this case the most probable number) is quite improbable. Clearly, a hypothesis cannot be rejected on this basis.

Now consider the case of 4 out of 20. An untrained person might be surprised at this figure and inclined to doubt that it had arisen by chance, and he would be right. He would also be equally or more surprised if there had been 5 left-handed, or 6, 7, 8, . . . , or 20. The *total* probability of these cases is only .017; so if an observer made the following rule, "I will reject the hypothesis every time 4 or more left-handed children occur in a group of 20," he might well be wrong, but if in fact the hypothesis were always correct, he would improperly reject it only 1.7 per cent of the time in the long run. Since no one can hope to be right all the time, most people would be satisfied with the situation and would conclude that either the children had something special about them, i.e., were not a random sample, or the population from which they were drawn

had a higher percentage of left-handedness than 5 per cent, or both.

The above rule is reasonable, but only because of considerations which have not been given. It is reasonable because there exists an alternative hypothesis: that the children were drawn from a population with a percentage of left-handedness greater than 5 per cent. Further, the set of possible results with 6 to 20 left-handed children forms the best critical set for this alternative.

Note that, in building the critical set for this case, one starts with 20 out of 20. Suppose that in actual fact the children came from a population with 10 per cent left-handedness. Then 20 out of 20 will still be extremely improbable; yet it is the first result which should be incorporated in the critical set . . . because 20 out of 20 has the largest ratio

$$\frac{\text{Prob. under the alternative hypothesis}}{\text{Prob. under the null hypothesis}}$$

After 20 out of 20 comes 19 out of 20, 18 out of 20, etc., until a critical set has been constructed which presents a risk α as large as the experimenter is willing to accept.

ESTIMATION

Suppose that the proportion p of "red" things in the parent population is not known but that a random sample with r red out of n has been drawn from the population. What can be inferred from this about the value of p? The method of confidence intervals has already been discussed . . . with the binomial distribution as an example. It will now be considered in more detail.

Equation (1) permits a direct calculation of the probability of drawing a random sample with r red out of n, when the proportion p in the whole population is known. Consequently, it can be used to construct a three-dimensional diagram in which the probability of r red out of n is plotted, for fixed n, as a function of both r and the proportion p. Figure 2 is such a diagram for $n = 20$. It will be seen that it consists of a ridge running from the lower left-hand corner ($p = 0$, $r = 0$) to the upper right-hand corner ($p = 1$, $r = n$). This ridge is highest (unity) at its two ends and sags to a height of about 0.177 in the center ($p = .5$, $r = .5n$). The probability falls off on either side of the ridge, rather steeply and unsymmetrically near the ends and less steeply in the middle.

The method of confidence intervals involves marking off on the pr plane an area on which rests some definite fraction (say 95 per cent) of the volume of the probability ridge. Such a division of the pr plane is shown in Fig. 3. . . . It has the property that the area between the curves contains the samples which will be drawn in 95 per cent of all trials, in the long run. Consequently the user will be wrong only 5 per cent of the time if he makes the assumption that a sample containing a known number of red items did actually belong to the inner area, *i.e.*, did come from a population with a proportion p lying in the range p_1 to p_2 thereby determined.

If the reader is entirely convinced by the above arguments, he has not fully absorbed the ideas of . . . [the first section]. Consider for example Fig. 4. This has been constructed with

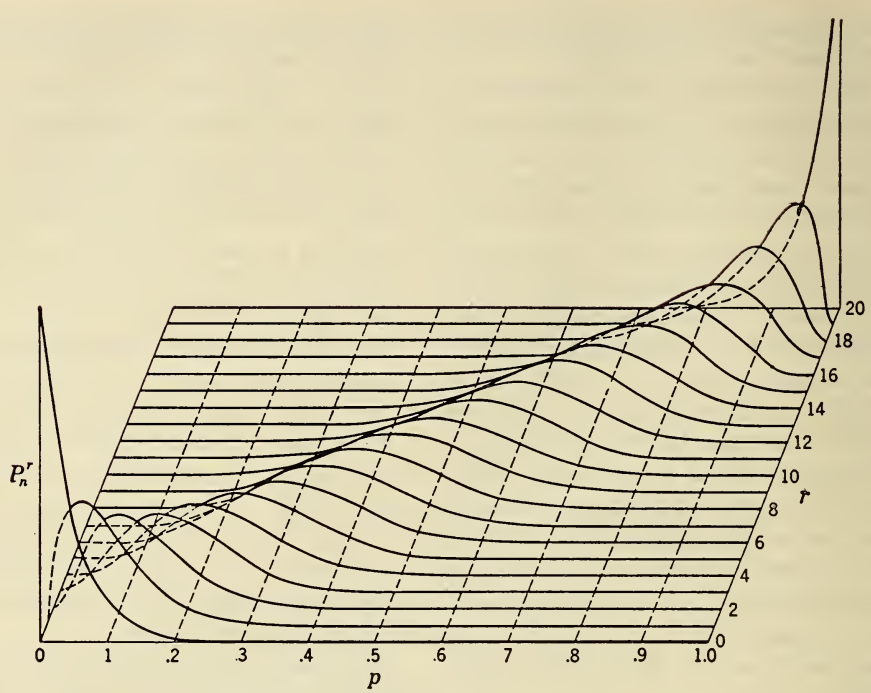


FIGURE 2

Diagram showing the probability P_n^r of drawing r, marked items in a random sample of 20 from a population with a fraction p of marked items. $n = 20$.

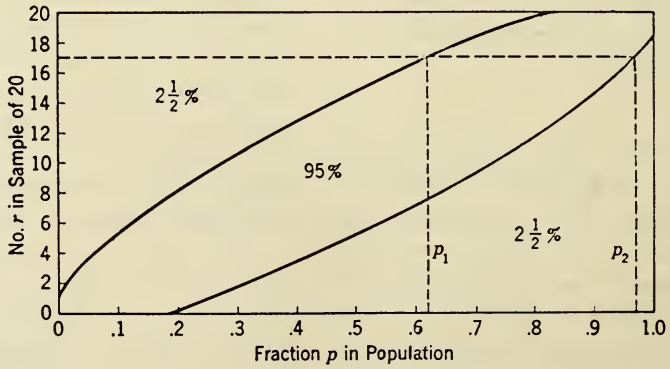


FIGURE 3

Ninety-five per cent confidence-interval diagram, for binomial distribution, $n = 20$. Area between curves contains 95 per cent of all possible random samples of 20.

the use of Eq. (1) so that the area within the curves has 95 per cent of the total probability, as before, but for each value of p only 1 per cent or fewer samples have values of r exceeding the upper limits and less than 4 per cent have values of r less than the lower limit. This figure can thus be used to construct confidence intervals also. For example, if r is observed to be 17, then the assumption that the sample came from the inner area limits the population value p to the range .57 to .96. But from Fig. 3 this sample example gives the range .62 to .97. Why are the two ranges different, and which is the better? In both cases the risk that the interval does not cover the true value of p is the same, namely, 5 per cent.

The situation here is similar to that of . . . [the first section]. The 5 per cent is the risk of not covering the true value, but there are many ways in which the rp plane could be marked off so as to yield this same risk. Clearly other considerations must be brought to bear. The usual one is the desire to make the confidence interval as short as possible. If this is used, the symmetrical scheme of Fig. 3 will be employed. . . . On the other hand, the penalty for overestimating p might be more severe than that for underestimating it so that an unsymmetrical scheme might be required. Note that in this example the difference is not very great between the two figures. With other situations, including various kinds of prior knowledge, it might be logical to mark off the area of the rp plane in other ways.

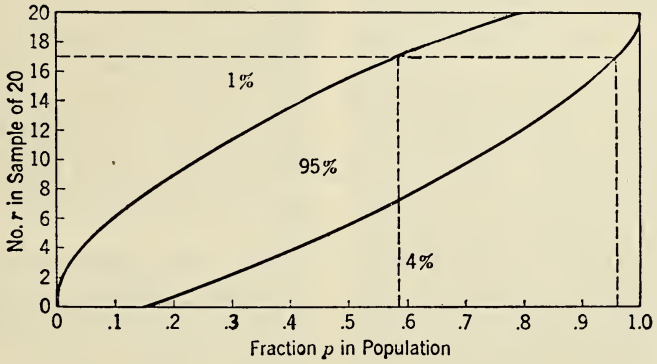


FIGURE 4

Unsymmetrical confidence intervals for binomial distribution. Ninety-five per cent of all samples lie between curves, but only 1 per cent lie above and 4 per cent below. $n = 20$.

III MONTE CARLO METHOD: SIMULATED SAMPLING

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE MONTE CARLO METHOD

DANIEL D. MC CRACKEN

During World War II physicists at the Los Alamos Scientific Laboratory came to a knotty problem on the behavior of neutrons. How far would neutrons travel through various materials? The question had a vital bearing on shielding and other practical considerations. But it was an extremely complicated one to answer. To explore it by experimental trial and error would have been expensive, time-consuming and hazardous. On the other hand, the problem seemed beyond the reach of theoretical calculations. The physicists had most of the necessary basic data: they knew the average distance a neutron of a given speed would travel in a given substance before it collided with an atomic nucleus, what the probabilities were that the neutron would bounce off instead of being absorbed by the nucleus, how much energy the neutron was likely to lose after a given collision, and so on. However, to sum all this up in a practicable formula for predicting the outcome of a whole sequence of such events was impossible.

At this crisis the mathematicians John von Neumann and Stanislas Ulam cut the Gordian knot with a remarkably simple stroke. They suggested a solution which in effect amounts to submitting the problem to a roulette wheel. Step by step the probabilities of the separate events are merged into a composite picture which gives an approximate but workable answer to the problem.

The mathematical technique von Neumann and Ulam applied had been known for many years. When it was revived for the secret work of Los Alamos, von Neumann gave it the code name "Monte Carlo." The Monte Carlo method was so successful on neutron diffusion problems that its popularity later spread. It is now being used in various fields, notably in operations research.

To illustrate the method let us start with the simple, classic Buffon needle problem. You get a short needle, draw on a sheet of paper several parallel lines spaced precisely twice the length of the needle apart, and then toss the needle onto the paper again and again in a random fashion. How often will

Scientific American, *May 1955, 192:5, 90–95.*

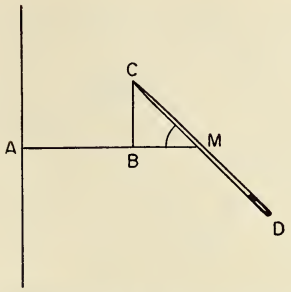
the needle land on a line? The mathematicians say that the ratio of hits to trials should be 1 to 3.1416. That is, dividing the number of hits into the number of throws, you should come out with the number 3.1416 (pi) if you continue the trials long enough (and throw the needle truly at random, without trying either to hit or to miss the lines).

I tried the experiment, with the following results. In the first 10 throws, the needle landed on a line four times. In the language of the statistician, there were four "successes" in 10 trials. The quotient is 2.5, which one must admit is not very close to 3.1416. In 100 trials there were 28 hits for an estimate of 3.57, also not good, but better. After 1,000 trials there were 333 hits for an estimate of 3, and my arm was tired.

This was hardly good enough to quit on, but the improvement with increasing numbers was not rapid, so it did not seem practicable to go on by hand. The fact is that the accuracy of a Monte Carlo approximation improves only as the square of the number of trials: to double the expected accuracy of the answer, you must quadruple the number of trials. I decided to make a calculating machine do the work, and I translated the problem to a medium-sized electronic calculator.

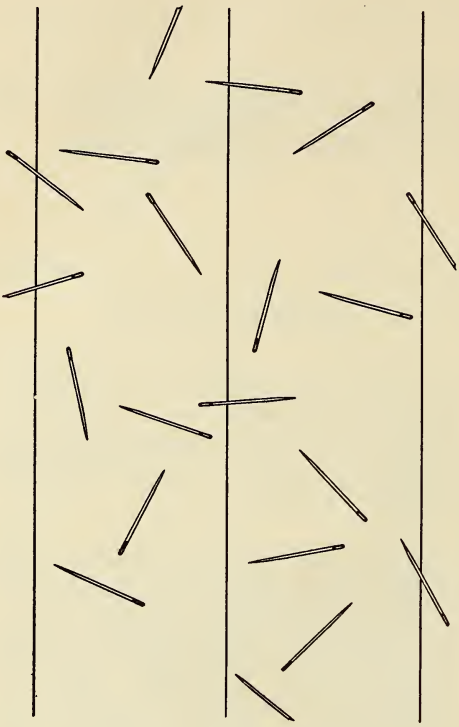
It is no difficult matter to make a calculating machine carry out operations which simulate the results of dropping a needle on ruled paper. Consider the diagram . . . [which appears below]. To describe the situation to the machine we must decide on a way of specifying the position of the needle relative to the nearest line. It does not matter on which side of this line the needle lies; nothing is changed if we turn the paper around. We can see that the distance from the midpoint of the needle to the nearest line (MA) is specified by a number between zero and two inches. The only other information needed to specify the position of the needle completely is the angle it makes with the perpendicular (MA) to the line. The angle is somewhere between zero and 90 degrees (not 180 degrees, because we are concerned only with the closer end of the needle). Given these two quantities, the machine can easily decide whether the needle touches a line; all it needs to do is to compute the distance MB (the cosine of the angle) and note whether it is less or greater than the distance MA—in the machine's terms, whether the difference is positive or negative.

Now to find out by experiment in what proportion of the trials a needle dropped at random would touch the



NEEDLE PROBLEM

is illustrated by a needle lying on a piece of paper ruled with parallel lines. The length of the needle is two inches; the distance between the lines, four inches. If the needle is thrown on the paper at random, how often will it land on one of the lines?

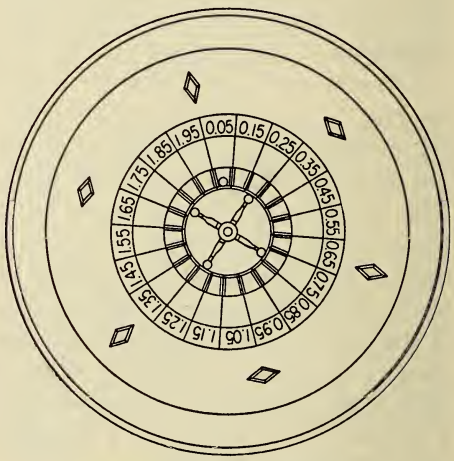


ACTUAL EXPERIMENT

on the needle problem was tried by artist Eric Mose. Each needle represents a toss and shows the position in which the needle landed with respect to the lines. In a sufficiently large number of trials, the ratio of hits to trials will be 1 to 3.1416 or pi.

line, we would like to test all possible positions in which the needle might land. To do this we would have to consider all possible combinations of distances and angles—essentially the method of the integral calculus. Obviously we are not going to tackle this infinite task. But in place of attempting a systematic exploration of all positions, we can take a random sample of them, and this should give us a reasonably accurate approximation of the correct answer, as a sampling poll may do.

How shall we select the random sample? This is where the Monte Carlo method comes in. Suppose we built a roulette wheel with 20 compartments, representing 20 different distances from the line (up to two inches) for the needle midpoint. A spin of the wheel would select the distance for us in a random manner, and over many trials each of the 20 distances would be selected about the same number of times. With a similar wheel we would pick the angle each time in the same random fashion. Then a series of spins of the two wheels would give us a random set of positions, just as if we had actually dropped a needle at random on ruled paper.



ROULETTE WHEEL

especially designed for the needle problem depicted . . . illustrates a basic feature of the Monte Carlo method. Each compartment of the wheel represents one of 20 distances between zero and two inches, the length of the needle.

Of course the wheel-spinning method would be more cumbersome than dropping the needle, but there are ways of

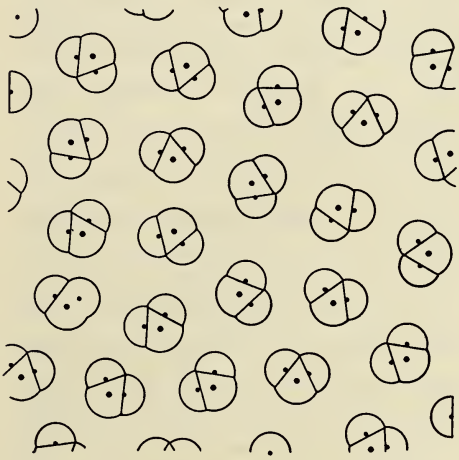
doing about the same thing with numbers and a calculating machine. First we get up two lists of numbers: one for distances in the range between zero and two inches, the other for angles in the range between zero and 90 degrees. The numbers are chosen at random to cover the whole range in each case without favoring any part of the range; we can take them from some list of numbers already checked for randomness or we can make our own list from, say, a table of logarithms, taking the numbers' last three digits. Then we put the calculator to work computing whether various combinations of the distance and angle numbers place the needle on a line or not (i.e., whether the difference between MB and MA is positive or negative). Repeating the operation many, many times, we can get as close to precision as we like; statistical principles tell us the degree of precision we can expect from a given number of trials.

The moderately fast computer I had available when I made the experiment was able to perform 100 "trials" per minute. In about an hour the machine ran through 6,000 trials, and there were 1,925 "hits." In other words, the estimate of pi was 3.12, which is as good as can be expected for 6,000 trials.

Even this simple case required a rapid computer. Most applications of the Monte Carlo method of course are much more complex. However, the present high-speed computers make them feasible: there are machines which can perform 5,000 trials per minute on the Buffon needle problem.

Let us see now how the method works on a simple problem in neutron diffusion. Suppose we want to know what percentage of the neutrons in a given beam would get through a tank of water of a given size without being absorbed or losing most of their speed. No formula could describe precisely the fate of all the neutrons. The Monte Carlo approach consists in pretending to trace the "life histories" of a large sample of neutrons in the beam. We imagine the neutrons wandering about in the water and colliding occasionally with a hydrogen or oxygen nucleus—remember that to a neutron water looks like vast open spaces dotted here and there with tiny nuclei. We shall follow our neutrons one by one through their adventures.

We know how far a neutron travels, on the average, before it encounters a



NEUTRONS wander through water in a series of events, each with a known probability. Here the microscopic structure of water is depicted in highly idealized form as consisting of simple molecules of H_2O. The larger sphere in each molecule is oxygen; the two smaller spheres are hydrogen. The neutrons may be absorbed by either an oxygen or a hydrogen nucleus or may bounce off from the collision. Some may escape from the water.

nucleus, the relative probability that this encounter will be with oxygen or with hydrogen, the relative chances that the neutron will be absorbed by the nucleus or bounce off, and certain other necessary information. Let us, then, take a specific neutron and follow its life history. It is a slow-moving neutron, and its first incident is a collision with a hydrogen nucleus. We know (from experiments) that the chances are 100 to one the neutron will bounce off from such a collision. To decide what it will do in this instance, we figuratively spin a roulette wheel with 100 equal compartments marked "bounced off" and one marked "absorbed." If the wheel says "absorbed," that is the end of the neutron's history. If it says "bounced off," we perhaps spin another appropriately marked wheel to decide what the neutron's new direction is and how much energy it lost. Then we must spin another wheel to decide how far it travels to the next collision and whether that collision is with oxygen or hydrogen. Thus we follow the neutron until it is absorbed, loses so much energy that it is no longer of interest or gets out of the tank. We go on to accumulate a large number of such histories and obtain a more or less precise figure for the percentage of neutrons that would escape from the tank. The degree of precision depends on the number of trials.

In practice, of course, we do not use roulette wheels but random numbers, as in the previous example. I have omitted much of the detail of the calculation for the sake of simplicity and clarity. In one very simple problem on which I assisted, an electronic calculator labored for three hours to trace the life histories of 10,000 neutrons through 1.5 million collisions. I would have had to sit at a desk calculator for some years to accomplish the same results.

As a third illustration of the Monte Carlo method, let us take a simple problem in operations research. Imagine a woodworking shop consisting of a lathe, a drill press and a saw, with three men to operate the machines. The shop makes one model of chair and one model of table. The question is: How should the work of the shop be scheduled to yield the greatest production, considering a number of variable conditions affecting output?

Certain basic information must be gathered before any calculation can begin. How long does it take on each machine to do the necessary work on each piece of wood? How much does the time needed for each job fluctuate because of fatigue, boredom or other personal factors? How frequently do the machines break down? After the data are gathered, a way is devised to make the computer simulate the operation of the shop under specified conditions of scheduling. We will not go into the details here; perhaps enough has been presented in the other examples to give an indication of what has to be done. The computation is properly classified as Monte Carlo because it is necessary to spin a roulette wheel, or the equivalent, to pick samples from the known distributions. For example, we may know that a certain job may take anywhere from 12 to 16 minutes, and we have noted the percentages of the cases in which it is performed in 12, 13, 14, 15 and 16 min-

utes respectively. Which time shall we use for a particular case as we follow the course of a day's work in the shop? The question must be decided by random sampling of the type I have described.

With the Monte Carlo method high-speed computers can answer such questions as these: How should the schedule be changed to accommodate a market change demanding twice as many chairs as tables? How much could the shop produce, and at what cost, if one man should be absent for two days? How much would the total output be increased if one man should increase his work rate 20 per cent? Under a given schedule of work flow, what percentage of the time are the men idle because the work is piled up behind a bottleneck machine? If money values can be assigned to idle time, loss of orders due to low production and so on, dollars-and-cents answers can be given to problems of this kind in business operation.

The Monte Carlo method, in general, is used to solve problems which depend in some important way upon probability—problems where physical experimentation is impracticable and the creation of an exact formula is impossible. Often the process we wish to study consists of a long sequence of steps, each of which involves probability, as for instance the travels of the neutron through matter. We can write mathematical formulas for the probabilities at each collision, but we are often not able to write anything useful for the probabilities for the entire sequence.

Essentially the Monte Carlo method goes back to probability theory, which was developed from studies of gambling games. But it takes the opposite approach. The mathematicians who originated the probability theory derived their equations from theoretical questions based on the phenomenon of chance; the Monte Carlo method tries to use probability to find an answer to a physical question often having no relation to probability.

In the neutron problem, for example, the investigator's thinking might have been along these lines: "I have a physical situation which I wish to study. I don't think I'll even try to find an equation representing the entire problem. Even if I could find one, which is very doubtful, I probably wouldn't be able to get much useful information out of it. I'll just see if I can't find a game of chance which will give an answer to my questions, without ever going through the step of deriving an equation." In some other situations the investigator would reason: "The physical situation I am interested in has resulted in an equation which is very difficult to solve. I cannot possibly solve it in any reasonable length of time by usual methods. I wonder if I could devise some statistical method which would approximate the answer to my problem."

Much work remains to be done on the method. One is always faced with the unhappy choice of either inaccurate results or very large amounts of calculation. A problem which demands 100 million trials of some "experiment" is still impracticable, even on the fastest present computers. Another difficulty is that it is seldom possible to

extend the results of a Monte Carlo calculation to another set of conditions. For instance, after we have solved the problem of the passage of neutrons through ordinary water, we have to start all over again to find out how they will behave in heavy water. Nevertheless, in spite of its various limitations the Monte Carlo method is able to give at least approximate answers to many questions where other mathematical techniques fail.

Many mathematicians are working to improve the method, especially to reduce the computation required and to determine exactly how much reliability can be attributed to its results in various types of problems. Up to now the technique has been used mainly on problems of nuclear physics, such as the diffusion of neutrons, the absorption of gamma rays, atomic pile shielding and the like. In the author's opinion, one of the most promising applications of the method is in operations research. It could be useful not only on production problems such as the one described here but also in telephone operation, traffic control, department-store inventory control and so on. Some of these possibilities are already being investigated. It is safe to say that we shall hear more from Monte Carlo in the next few years.

✦✦✦✦✦✦✦✦✦✦✦✦✦ PREDICTION OF CONSEQUENCES: *Monte Carlo techniques*

Before we can hope to find out the consequences of a course of action, it is necessary to have some sort of data.

For example, suppose it is desired to test a proposed inventory policy to see how often the product will be out of stock, and how great the volume will be of deliveries which are delayed because there is not enough inventory on hand to meet customer orders. These figures, of course, depend on the fluctuations in the level of customer demands—how frequently total demand will exceed a particular magnitude.

In solving such a problem, experimentation is not a real possibility. It can be costly in customer relations to try very much lower inventory levels just to see what will happen. Moreover, the experiment will add little more to our knowledge (aside from offering us data on just how angry customers do get) than does simply waiting to see what demands turn out to be with current inventory levels. Past experience can be used to construct an imaginary experiment—to see what would have happened if the proposed inventory level had been used in the past. But if the product is relatively new, this can provide only very limited

Cost and Profit Outlook, *April 1957, 10:4, 2, with permission of Alderson Associates, Inc., Marketing and Management Counsel.*

information; and waiting around for experience to accumulate may be very time consuming and (if the current choice of inventory level is poor) very expensive.

The operations researcher has, however, invented another very effective way to gather the relevant data: that is, to make them up himself, or rather to let the mathematical statistician make them up for him! Like the cutting of the Gordian knot, this may strike the reader as a direct and ingenious approach, but one which does not meet his original conception of the problem. How can improvised statistics help us to foresee what will happen in the real world?

The answer is that the numbers are invented in a manner which carefully employs the analytical methods of mathematical statistics in order to stretch as far as possible such few actual data as are available to begin with. In the absence of any change in any of the major purchase influences, such as a sharp temporary price cut (a sale) or an advertising campaign, we may assume that customers will arrive randomly, in a pattern somewhat similar to outcomes in successive throws of a pair of dice. The pattern of customer demands may then be described in terms of a frequency distribution, which indicates how many weeks in a year customer demand can be expected to fall between 50,000 and 55,000 units, how often the demand will lie in the 55,000 to 60,000 range, etc.

Now, from the available information *and the nature of the problem,* the statistician can decide which frequency distribution best describes the pattern of expected customer demands. From the frequency distribution it is then possible to construct an artificial history of customer demands by choosing randomly among all the possibilities, but in a way which is "loaded" to produce the right frequencies. To give a very simple illustration, suppose we consider two possibilities: *A,* weekly demand less than 50,000 and, *B,* weekly demand of at least 50,000. If, on some basis, the odds are computed to be 2 to 1 in favor of *A,* we can generate an artificial demand history as follows: Toss the (unbiased) die. If it falls 1, 2, 3, or 4, put down an *A;* whereas if it falls 5 or 6, put down a *B.* This might yield a pattern for weekly demands such as the following:

		Sales "History"	
"Week"	*Face of Die*	*Under 50,000*	*50,000 or More*
First	3	*A*	
Second	1	*A*	
Third	3	*A*	
Fourth	5		*B*
Fifth	6		*B*
Sixth	2	*A*	
Seventh	2	*A*	

This, incidentally, indicates the reason for the term "Monte Carlo method."

In practice, it is not actually necessary to toss any dice. Instead, we can use tables called "tables of random numbers" which have been worked out in advance. Moreover, the computations can be made by high-speed electronic computers which are able, in a few minutes or hours, to run off thousands of cases and manufacture data

whose collection would, otherwise, require many years. But although this method is economical and powerful, it must be used only with the greatest care and caution. Everything, as we have seen, depends on the choice of frequency distribution (i.e., the odds of the various outcomes), and unless there is some assurance that these have been picked well, the entire calculation can be worthless.

✦✦✦✦✦✦✦✦✦✦✦✦ NEW METHOD PRE-TESTS IDEAS

DONALD G. MALCOLM

MAKING A SIMULATION: HOW MANY REPAIRMEN ARE ENOUGH?

System simulation generally employs what is called the Monte Carlo method, so called because it permits study of situations which are as unpredictable as the fall of dice or the number which comes up on a roulette wheel. A simplified example from the production side of a business will show how . . . [the method works.]

A shop has 20 machines which run continuously for 24 hours per day. These machines break down from time to time despite preventive maintenance practices. Four repairmen are on duty during each of the three shifts and the machines are such that only one repairman can work effectively on them at a time. There also is a fixed production requirement for each week. This makes it necessary to work overtime on Saturday to make up for any down time during the week.

Sometimes more than four machines are down simultaneously while at other times all repairmen are idle. Even more frequently three and two men are idle. The question is: Are four repairmen the proper number?

An answer could be provided by experimenting with the real men and machines. An additional man could be put on, and it would be possible to measure how much better machine utilization would be achieved. Then this could be weighed against the cost of the extra repair help, and a measure of improvement reflected. After a while, the repair crew could be cut back to three men and the effects again be measured. However, this would be costly both in time and in money. System simulation is a way of performing this experiment in the laboratory. A simulation experiment using three repairmen requires some data already available from records concerning the running time between breakdowns and the elapsed time in making the necessary repairs. First, the operations ana-

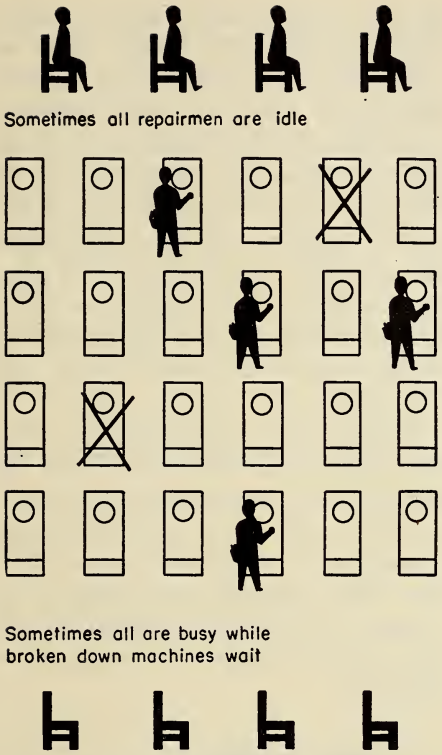


FIGURE 1

lyst advises there is one chance in 10 that a machine will break down in any given hour. Second, study of the necessary repairs discloses that 40 per cent of breakdowns require a half hour to repair; 30 per cent require an hour and so on. This data permits simulation of what would happen if various numbers of repairmen were used. Let's say the manager starts with three.

To estimate how many machines will break down in a given hour of operation of the simulated system he refers to a table of random numbers that have been recorded by scientists for just such types of analysis. These in effect may be thought of as the record of a million throws at a wheel. Since there are 20 machines, each with a chance for a breakdown, he looks at 20 numbers from this table, and arbitrarily says that number nine represents a breakdown. Note that there is only one chance in 10 of the number nine being read for each number looked at.

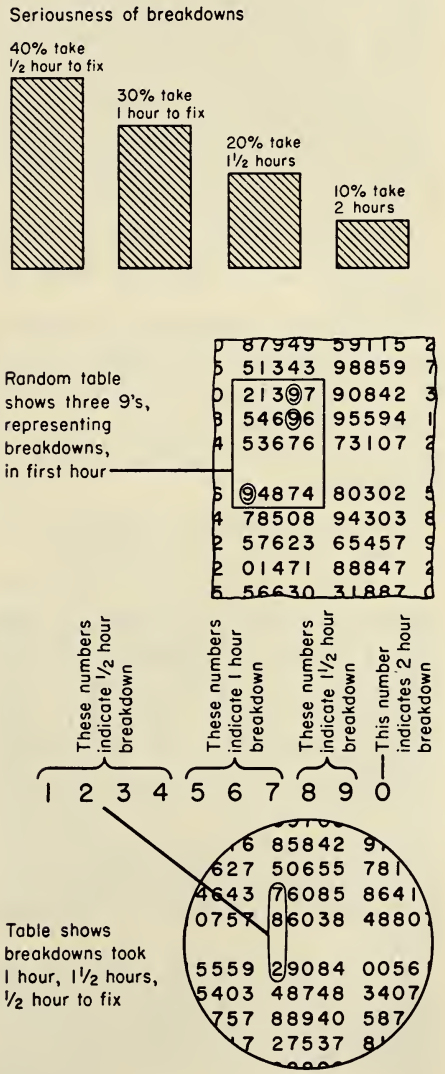


FIGURE 2

If three number nine's appear among the 20 numbers, he had three breakdowns in the first hour.

The next question is, "How long do repairs take?"

Using the random tables again, he chooses numbers to represent the varying lengths of repair time the different types of breakdown require. Since he already knows that repairs requiring a half-hour to fix occur, say, four times as often as repairs requiring two hours and twice as often as repairs requiring an hour and a half, he selects numbers in this same proportion.

These three repairs might be found to require 1, 1½, and 1/2 hours respectively.

He has now simulated one hour of operation of the man-machine system. He knows how many machines broke down in the first hour and how many man-hours of work will be required to repair them.

This same method is continued until many such hours are simulated.

As the simulation goes along troublesome situations will arise. In some time periods the manager will find that more work arrives than he can conveniently handle with his work force. He records the waiting time that is necessary due to this unlucky work load. This represents time which the machine was down and could not be worked on because the repairmen were all occupied on jobs that had arisen earlier.

At this point he has simulated the operations of this greatly simplified management system in the laboratory. This has been done without disrupting the organization by adding or removing one employe and without waiting until working habits settle down so that a real live study would produce meaningful data. He next performs simulations for two, four and five repairmen. Examination of these data will indicate how much he will gain in machine utilization through employment of additional repairmen. At this point his simulation is completed and he is in a position to put costs on the data in such a way as to make a lowest cost solution.

Suppose that, considering the overtime and standby labor involved, the manager has ascertained that idle machine time costs $5 an hour. The regular repairman rate is $2.50 an hour. Now with data available the manager can compute the cost per machine hour of operating with different numbers of repairmen.

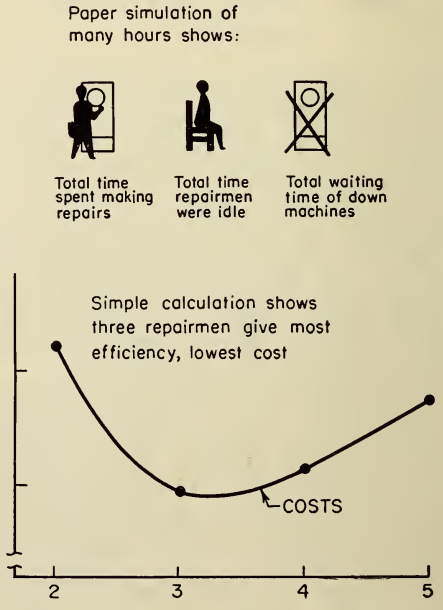


FIGURE 3

He can decide what number of repairmen will give him the most efficient operation and in this case his decision is indicated quite clearly—he should change to three repairmen instead of the four currently used. He can also estimate the annual savings which the most efficient number will give him.

At this point, the manager of the affected department may have some ideas concerning how he could get better results than indicated. He might, for example, suggest that giving priority to certain repairs could improve the machine utilization, or that purchase of more reliable machines would be a better course. System simulation can easily be extended to permit this manager to try out these ideas.

Generally speaking, most big problems worth exploring in great detail will be most efficiently and economically handled through use of an appropriate electronic computer. Even such a problem as the example above, when extended to include other conditions, can be explored more efficiently and quickly if performed on an electronic computer.

IV SIMULATION

✦✦✦✦✦✦✦✦✦✦✦✦ SYSTEM SIMULATION

DONALD G. MALCOLM

During recent years a new technique has come into prominence as an aid in both the task of training and in problem solving. This technique, called *System Simulation*, has been developed in both the military and industry by operations research and Industrial Engineering groups charged with making recommendations concerning complex planning problems and in developing training methods. A growing need for this thorough and scientific study, plus the availability of high speed electronic computers, has brought the concept of system simulation to the fore as a most useful Industrial Engineering and management tool.

System simulation has the most useful property of permitting experimentation with and testing of certain policy, procedure and organization changes in

Reprinted from the May–June 1958 issue of The Journal of Industrial Engineering, *9:3, 177–186, Official Publication of the American Institute of Industrial Engineers, Inc., 345 East Forty-Seventh Street, New York 17, New York. The author, Donald G. Malcolm, is the President of Management Technology, Inc.*

much the same way as the aeronautical engineer tests his design ideas in the laboratory or the "wind tunnel." Simulation, long used as an engineering method, is now being used to:

1. Study complex operating plans, and management controlling systems for the purpose of designing better plans and/or systems.
2. Study and train people in the operation of complex tasks.
3. Gain acceptance of proposed changes through better understanding of how a given system works or operates.

Simulation has application from very small day-to-day problems to complex management and industrial engineering problems requiring operations research teams and computers. As we shall see, simulation has the advantage of being easily understood, of being relatively free of mathematics and of often being quite superior to mathematical methods which may be too complex to apply or even not available. Another distinct advantage lies in the fact that simulation generally eliminates the need for costly trial and error methods of trying out a new operating concept on the real flesh and blood and machines. . . .

One of the difficulties involved in the solution of problems of the broad nature referred to, is the compartmentalized thinking that organization structure and accounting methods tend to impose upon the managers of the affected departments. Each manager is motivated primarily by goals related to his own function which generally conflict with cost goals in other departments. The management of inventory is a well-known example of this class of problems.

In solving such problems top management generally arrives at a solution by compromise—a little bit for everyone and really not getting the best economic solution from the over-all company point-of-view. Since this is a well-known problem, let us briefly discuss the question it raises. How can we get at such problems where interactions between various elements of the problem or between various functions in the organization play such an important role in finding this best overall solution, and, perhaps even more important, getting the solution accepted and implemented?

METHODS FOR ATTACKING SUCH PROBLEMS

There are three general courses of action open to us which are of value in different ways. Let us list these and then discuss each method briefly.

1. Experiment with the real facilities, machines and men.
2. Formal mathematical analysis—construct equations describing the various alternatives.
3. Conduct simulated experiments, or *System Simulation,* as herein termed.

EXPERIMENT WITH THE REAL FACILITIES, MACHINES AND MEN

Generally speaking, any idea that has been thoroughly studied is still not proven until it is tried out in the real situation. In using this method, suggested ideas or plans are brought forward for consideration as a result of cost studies of varying degrees of detail and refinement. Or, we may decide

to emulate the practices of successful competitors and plan to do better through knowledge of their successes and mistakes.

In either event the basic task of study and analysis of the interactions involved by means of "models" is not performed. The suggested plan, when approved, is simply put into effect. If the plan proves to be inadequate a modification is made. Thus, as often happens, we end up "experimenting" with our own men, machines and facilities. While in many cases this is the only method available—*and it does get action*—there is the attendant confusion and cost of operating inefficiently for a long period of time while "testing" alternatives that may be avoided by the analysis or simulation approaches.

FORMAL MATHEMATICAL ANALYSIS

This is the most desirable and powerful approach. However, this method, which may be said to consist of writing equations which describe completely the problem area or system under study, is often too complicated to utilize effectively. Also in many situations the mathematics have not or cannot be developed which will permit all the desired factors to be considered simultaneously. This is particularly true in the case of competitive problems.

Further, the mathematical method generally poses a distinct problem in communication. It is often hard to convince people that what a complex formula seems to say is really the best thing to do. There are many "if's" and "but's" that are hard to overcome in the process of getting change accepted. Thus "experimentation," as discussed above, generally still has to be performed. In cases where the first two methods have not given satisfactory results the method of *System Simulation* is often useful.

SIMULATED EXPERIMENTS—SYSTEM SIMULATION

In using this method the problem or system under study is first described as the sequence of individual operations to be performed. This may be called the "Flow Model" of the system. It is then necessary to have data indicating how the individual operations are interrelated and to have the frequency distribution of elapsed times for each individual operation for the different conditions to be explored.

Then inputs of such items as manpower, scheduling methods, or amounts of equipment, facilities, etc. are systematically varied. By consulting the time data mentioned above in a random manner, the over-all time for the sequence of operations can be determined. This process performed over and over simulates operation of the system and permits accruing such total system data as average equipment and manpower utilization, or inventory outages, delays etc. Such outputs are then used to evaluate the desirability of the given input under test and in effect a simulated experiment has thus been conducted. . . .

THE ROLE OF THE COMPUTER

Up to this point the use of a computer in connection with the concept of simulation has not been mentioned. Generally speaking, most problems of

any size worth exploring in any great detail will be most efficiently and economically handled through utilization of an appropriate electronic computer. The volume of detail and the number of hours that must be simulated in order to arrive at meaningful results gen-

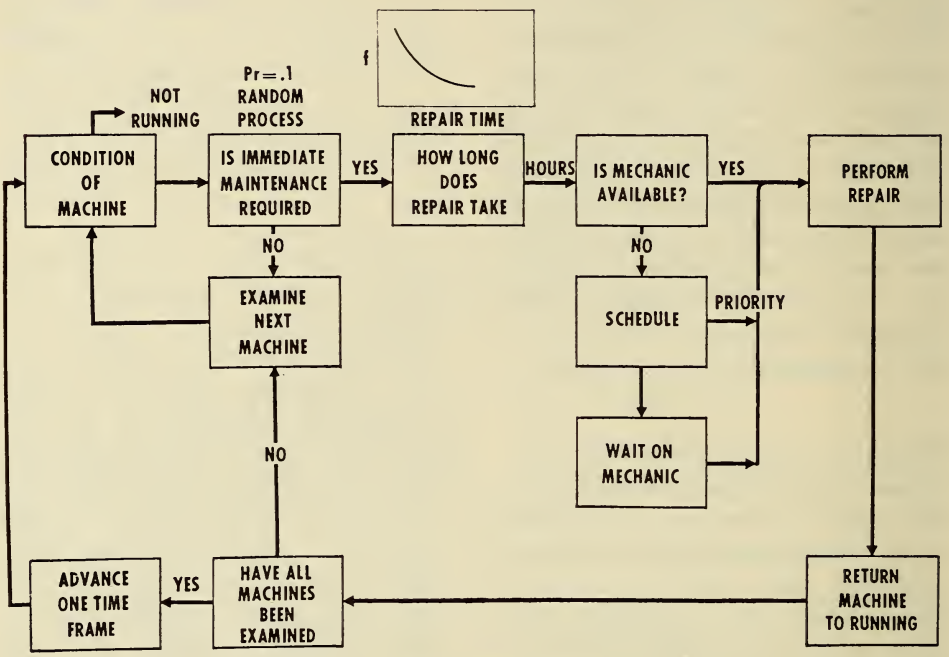


FIGURE 1

FLOW DIAGRAM FOR SYSTEM SIMULATION

erally demands that such a program be developed. . . .

In preparing the problem for the computer, the analyst generally prepares a logical Flow Diagram. This diagram shows the sequential steps involved in the problem or "system" under study. Fig 1. is typical of the kind of diagram that the analyst prepares in describing the system under study. This diagram is useful in the programming of the computer.

INDUSTRIAL APPLICATIONS

Perhaps the most dramatic aspects of system simulation lie in its ability to reproduce the workings of large scale systems. While such ambitious programs are of more interest, it should be borne in mind that their success is dependent on having performed many smaller simulations for experience, upon adequate data and upon consideration of the many mathematical and statistical problems that are involved in construction of a feasible and eco-

nomical model. Those considering broad full scale models as panaceas in the decision making area are well advised to engage in smaller projects at the outset. This will prevent certain disillusionment and make for a more meaningful research program.

With this word of warning let us turn our discussion to a few industrial applications of system simulation.

DISTRIBUTION AND INVENTORY CONTROL MODEL

Imperial Oil has studied an extensive distribution system composed of many hundred field warehouses in connection with an extensive expansion problem. The flow of stock under various possible combinations of facility plans was studied by simulation and costed. Results of this work were instrumental in suggesting a central warehousing method of operation.

Inventory and distribution problems are particularly complex and interrelated and are among the most important problems in management today. It is understandable, therefore, that we find considerable activity in the simulation of various distribution and inventory problems. . . .

AIRPORT STATION MODEL

United Air Lines has set up and operated a simulation of the operations at a large airport on an IBM 704 computer. With it, several months of actual operations at an air terminal can be simulated in a matter of minutes. If one visualizes the "flow" of a plane through an airport and all the factors that determine the total time spent, such as landing, taxiing, maintenance, loading, etc., an idea of the complexity of the simulation is afforded. There are reportedly some 9,000 instructions in the computer program which took some nine months to develop after the logic was worked out in flow form. A diagram for this simulation, similar to Fig. 1, would be quite complex!

This model involved such structural and environmental elements as:

1. Time of day, week, and year
2. Weather conditions
3. Maintenance plan
4. Availability of spare aircraft
5. Manpower schedule

Probabilistic factors such as the need for maintenance, the type and length of resulting repairs, absenteeism of personnel and scheduling delays form the variable background for testing changes in company policies and practices. For example, management can change the number of spare aircraft, or the manpower schedule, and simulate operations under these new conditions. The computer is programmed to provide such output measures of performance as expected idle manpower, expected idle equipment, utilization of maintenance and berthing facilities. By comparing the expected performance in terms of these outputs with the cost of obtaining that performance, decisions can be reached that produce the best over-all operation of the complex airport system.

PRODUCTION CONTROL MODEL

The General Electric Company is using simulation to test new concepts and methods in production scheduling. Every production control manager knows that the methods, priority rules

and procedures that he uses in discharging his scheduling job have an important effect on his company's utilization of men and machines and also upon how smoothly the production flows and serves the customer.

In this model, factory operations are simulated to test these decision rules which, in effect, are the scheduling system. Policies and procedures concerning machine loading, scheduling, and dispatching are being systematically tested in the laboratory and evaluated in terms of internal inventory cost, idle man and machine time, flexibility and cost of the scheduling itself. In this way the trial and error method of actually trying out a new approach will be avoided. . . .

PROFIT PLANNING SIMULATION

A television tube manufacturing company has explored ways of obtaining better profit from its over-all operations through simulation. The flow of product was studied throughout the entire system. A flow chart of material from the point of purchase through operations in manufacturing to transportation in the distribution system was made. This chart served as the basis for programming a computer. Such factors as changes in volume imposed by the changes in customer demand, changes in product mix from a scheduling point of view, changes in the number and location of manufacturing plants, changes in the method and pattern of distribution and storage, and different expected loses of product in the manufacturing process could be made in the program put into the computer.

By studying these changes, it was possible to find out which factors were of the greatest sensitivity in producing profit and it was possible to experiment with possible changes in any of the factors indicated. Finally, the company feels that it has a realistic mechanism for planning and controlling its business system.

A TOP MANAGEMENT DECISION SIMULATION

The American Management Association has had constructed a "game" or "competitive simulation" built around the major control decision problem that a top executive faces in piloting the course of his company in a competitive economy. This simulation exercise has been made the central subject of a two week course in Executive Decision-Making at Saranac Lake, N.Y.

In this exercise, executives must make decisions concerning the allocation of funds to production, marketing, research and development and new plant construction activities as well as set the price of their product and specify market research information to be procured and obsolete plant to be sold. These decisions must be made while in competition with four other companies, each striving for a greater return on their investment. The objective of this exercise is to give appreciation training to junior executives concerning the need for balance and the development of competitive strategies in plotting a company's course during ten years of simulated operations. Forms similar to the one shown . . . [here] are filled out every fifteen to twenty minutes (a quarter of a year's operation) and turned in for processing by an IBM

650 computer, which is programmed to determine the effect of their decision and print out the results.

Plotting of information, planning and budgeting activities and the critique session at the end of the play serve to bring out quite realistically many of the principles of scientific management.

COMMON THREADS IN THE USE OF SIMULATION

A survey of users of simulation gives us answers to two questions that are of interest. Why use simulation?—the purpose. Under what conditions is it used?—the type of problem.

The purposes may be summed up as follows:

1. To train people in their duties in a complex system.
2. To learn about the operation of a complex system. This may be for either research or training purposes.
3. To experiment with possible or suggested changes in policies, procedures, men, machines, under laboratory conditions and in a dynamic context.
4. To help develop a master plan for systematic management research leading to the development of new control systems.
5. To study the decision maker in a live situation.

The conditions under which simulation has been employed include the following:

1. Where uncertainty, or variability of data is an important factor and where direct mathematical methods are not available or practical to employ.
2. Where detailed procedures must be developed and alternatives can be evaluated only in terms of total system costs.
3. Where large, complex problems with many interrelated factors are to be studied and simpler models do not satisfactorily handle the interactions.
4. Where disagreement exists between affected parties in regard to proposed changes and an acceptable method of arbitration is required.

TRAINING THROUGH SIMULATION TECHNIQUES

In addition to the AMA simulation there are many other training simulations which have not been publicized as yet. Such companies as Westinghouse Electric Corporation, General Electric Co. and the Rand Corporation are using simulations, . . . to train their inventory people in problems of operating and providing satisfactory stocks in their distribution systems.

Results to date seem to indicate that simulation provides a means of more effective training and that it promises to augment and even replace some of the management training methods used in colleges. Therefore, perhaps a brief summary of some of its advantages is of interest.

ADVANTAGES OF TRAINING THROUGH SIMULATION

1. Permits "cause" and "effect" to be felt. The individual can appraise the effect of his actions immediately.
2. Aids in judging what information is really important in making decisions and obtaining management control.
3. Familiarizes the individual with the data needed and available to him in making his decisions, thereby demonstrating economic and management principles in practice.

TOP MANAGEMENT

STATEMENT OF ASSETS

YEAR 0 QUARTER 0

		TOTAL	NET CHANGE
CASH		$ 4,425,000	$25,000
INVENTORY	150,000 units @ $ 4.50	$ 675,000	$ 0
PLANT INVESTMENT	1,010,000 units @ $ 5.00	$ 5,050,000	$50,000
TOTAL ASSETS		$10,150,000	$75,000

INCOME STATEMENT

SALES INCOME	900,000 units @ $ 5.00		$4,500,000
COST OF GOODS SOLD & OPERATING EXPENSES			
COST OF GOODS SOLD		$4,050,000	
MARKETING & RESEARCH AND DEVELOPMENT		$ 300,000	
OTHER (MARKET RESEARCH)		$ 0	$4,350,000
			$ 150,000
OTHER INCOME (PLANT DISPOSAL)			$ 0
INCOME BEFORE TAXES			$ 150,000
TAXES			$ 75,000
NET INCOME			$ 75,000

OPERATING AND DECISION

(for next

UNIT COST OF PRODUCTION	$ 4.65	$ 4.61
UNITS OF PRODUCTION	720,000	765,000
		DECISION
COST OF PRODUCTION	$3,348,000	$3,526,700
MARKETING		$ 170,000
RESEARCH & DEVELOPMENT		$ 85,000
ADDITIONAL PLANT INVESTMENT	$ 0	$ (10,000)

MARKET RESEARCH INFORMATION

S=COMPETITORS' SHARE OF MARKET		S
M=TOTAL INDUSTRY MARKETING EXPENDITURE	NONE	$ 5,000
R=TOTAL INDUSTRY RES & DEVELOP EXPENDITURES		
A=POTENTIAL MARKET SHARE – MAX. MARKETING		
P=POTENTIAL MARKET SHARE – MAX. PRICE		
PRICE	$ 4.80	$ 4.85
PLANT DISPOSAL (in units)		
	1	2

IBM 650 REPORT

DECISION GAMING

ANNUAL STATEMENTS

YEAR 0

COMPANY 1	COMPANY 2	COMPANY 3	COMPANY 4	COMPANY 5
$ 4,425,000	$ 4,425,000	$ 4,425,000	$ 4,425,000	$ 4,425,000
$ 675,000	$ 675,000	$ 675,000	$ 675,000	$ 675,000
$ 5,050,000	$ 5,050,000	$ 5,050,000	$ 5,050,000	$ 5,050,000
$10,150,000	$10,150,000	$10,150,000	$10,150,000	$10,150,000

MARKET INFORMATION

	COMPANY 1	COMPANY 2	COMPANY 3	COMPANY 4	COMPANY 5
PRICE	$ 5.00	$ 5.00	$ 5.00	$ 5.00	$ 5.00
SHARE OF MARKET	20.00%	%	%	%	%
TOTAL MARKET	4,500,000				
POTENTIAL SALES	900,000				

MARKET RESEARCH REPORT

TOTAL INDUSTRY MARKETING EXPENDITURE $
TOTAL INDUSTRY RESEARCH & DEVELOPMENT EXPENDITURE $
POTENTIAL SHARE OF MARKET – MAXIMUM MARKETING %
POTENTIAL SHARE OF MARKET – MAXIMUM PRICE %

INFORMATION

period)

		DECISIONS LAST PERIOD				
$ 4.57	$ 4.54	$ 4.50	$ 4.49	$ 4.48	$ 4.46	$ 4.45
810,000	855,000	900,000	918,000	936,000	954,000	972,000
ALTERNATIVES						
$3,701,700	$3,881,700	($4,050,000)	$4,121,800	$4,193,300	$4,254,800	$4,325,400
$ 180,000	$ 190,000	$ 200,000	($ 210,000)	$ 220,000	$ 230,000	
$ 90,000	$ 95,000	$ (100,000)	$ 105,000	$ 110,000	$ 115,000	
$ 20,000	$ 30,000	$ 40,000	$ 50,000	$ 60,000	$ 70,000	$ 80,000
M	R	S & M	(S & R)	M & R	S, M & R	
$ 10,000	$ 10,000	$ 15,000	$ 15,000	$ 20,000	$ 25,000	
	A	P	A & P			
NONE	$ (22,500)	$ 22,500	$ 45,000			
$ 4.90	$ 4.95	$ 5.00	$ 5.05	$ (5.10)	$ 5.15	$ 5.20
	(NONE)	5,000	10,000			
3	4	5	6	7	8	9

TOTAL FUNDS AVAILABLE → $4,425,000

COMPANY 1 PERIOD 1 GAME 1

4. Is a means of providing some familiarization with Electronic Data Processing. This is an extra benefit to many participants who are not already familiar with computers.

5. Decisions must be made from facts available and analyzed immediately. Simulation poses a realistic situation, not an academic one, full of exceptions and sometimes too inconclusive in nature to be satisfying.

6. Because it is more realistic, the students or players get caught up in it and work more intently at the tasks it requires.

The success and utility of a simulation which is aimed at training depends on obtaining "involvement" of the participants, how adequately it represents reality and whether there is a transfer of skill back into the real life situation. At this point in the development of simulation, it is difficult to *prove* scientifically that there is a *real* transfer or that it *is* beneficial to the student.

However, everyone associated with simulation believes that an intelligent individual will not carry away specific acts to perform (unless so directed), but rather will carry away a point of view in tackling problems, thinking more of the system in which they reside, what courses of action are available, and what facts must be considered. In order to test the nature of what the student retains from his simulation exercise, AMA is having two psychologists build up scales to measure the quality and type of training received.

Thus, simulation training methods also give us the possibility of testing the effect of many training principles, and should prove a most versatile addition to our educational process. . . .

RELATION TO INSTALLATION THEORY

In the present world, change in the form of new products and processes is coming at a faster rate than ever before. The life of a new product or process is shorter. All of this adds up to the need for introducing change more efficiently and accurately since the learning time is now a much more significant portion of the product's total life span and cost. While system simulation is no panacea, it offers assistance and is a useful tool in such *installation theory.* As we have seen, it has utility in analyzing complex problems and in training operating management in the system it operates. Companies that clearly recognize that efficiency comes from the real understanding of broad operating objectives on the part of all levels of management involved, will clearly be more flexible and achieve lower cost operations more quickly.

The evidence is clear and encouraging that system simulation is a useful and powerful tool in helping to reduce installation time, in improving the quality of understanding and in searching for system cost reduction opportunities. In this technique would seem to lie the method for research into systems design.

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE USE OF SIMULATION IN MANAGEMENT ANALYSIS: *a survey*

DONALD G. MALCOLM

The purpose of these remarks is to review briefly what has been done and why in the area of computer simulation applied to management analysis and training. The role of the modern industrial manager is indeed a complex, perplexing one. For, operating in the face of uncertainty in our dynamic, competitive, nonlinear and probabilistic world, he is called upon to make major innovations and to adjudicate the proposals for change that come from the various functions or divisions of his total organization. This problem of obtaining economic balance, or least-total-cost operation as it may also be phrased, today, as well as in the plans for future operations, is most important to the success of his company. The need for such balance promises to become even greater in future years as we learn better to use the information-handling and decision-making apparatus made possible through electronic computers. As many writers have indicated, such successful use of computers may be a necessary ingredient to corporate survival.

Computer Simulation, or *System Simulation*, is proving to be a useful tool in this quest for better understanding and control of business operations. It is our purpose here to explore briefly what simulation is, who is doing it, for what purpose, and to speculate on its future outlook.

In this discussion, we shall define a simulation project as one which depicts the workings of a large-scale system of men, machines, material and information operating over a period of time in a simulated environment representative of the actual real world conditions. A simulation project generally utilizes an electronic computer and may operate in "real" time or in compressed time where the simulated system may be operated for several years in but a few hours of computer time.

In effect, the arrival of simulation on the management analysis scene is dependent on three concurrent developments. We have already touched on the first and perhaps most important, the need and desire on the part of management to have a more complete and accurate analysis of alternate courses of action and to obtain economic balance

Report of the Second System Simulation Symposium, American Institute of Industrial Engineers, Inc., 1960, 11–26. The author, Donald G. Malcolm, is the President of Management Technology, Inc.

in operations. Little more needs to be said here concerning this point.

The second development stems from the use of engineering systems analysis and/or operations analysis in this management area. Operations researchers have found that the number of variables, the probabilistic aspects, and the servo aspects of systems under study often defy the use of strictly analytical or mathematical methods of analysis.[1] Simulation of the system or process under study often permits sufficient alternatives to be evaluated, thus leading to a better, if not optimum, solution.

The third development making the use of simulation a practical tool is the advent of the high-speed electronic computer. Our discussion will be confined to the use of digital computers since they have been found, by and large, most applicable to the class of problems involved.

Since it has often been predicted that computers will become automatic decision-making devices and indeed many have been purchased with this stimulating and challenging goal in mind, it is well to spend a few paragraphs in describing the current state and categories of usage of computers in industry. In this way, the role of simulation can be put in better perspective and the specter of technological unemployment among decision makers (if it exists) can be dispelled.

[1] F. M. Ricciardi, C. J. Craft, D. G. Malcolm, R. Bellman, et al., *Top Management Decision Simulation*, American Management Association, 1957, and C. J. Thomas and W. L. Deemer, Jr., "The Role of Operational Gaming in Operations Research," *Operations Research*, 5:1–27, February 1957.

I. CATEGORIES OF USES OF COMPUTERS IN THE MANAGEMENT PROCESS

In making this categorization, we are delimiting our scope from the scientific and physical engineering applications of computers to those applications more generally thought of as involving and directly assisting the management of the organization. This area of application we have termed the "management process." Here, there are three major categories of uses to which computers have been put.

(1) A Data-Processing Tool
(2) A Problem-Solving Tool
(3) A Controlling Device

1. A DATA-PROCESSING TOOL

Since the first installation of an electronic computer for strictly business use at General Electric in Louisville in January 1954, there has been a mushroom-like growth in the application of computers to the routine automation of existing information, communication, personnel, and other data-reporting systems. This includes such applications as payroll, inventory position and control, production release and invoicing systems, etc. Of the $631 million of computers to be installed by June 1959, and the $591 million on order, it has been estimated that up to 90% of machine time will be used for applications in this category.

It should be pointed out that most of these systems, which are being automated, are not being evaluated in the sense that their ability to best perform the function for which they were de-

signed *is itself* critically examined. Most installations are undertaken for the purpose of reducing time in report preparation or in the attempt to effect clerical savings.

2. A PROBLEM-SOLVING TOOL

The computer is often used as the vehicle for research in connection with the design of management control systems for policy determination and for training purposes.

There is a considerable amount of simulation activity in connection with such problem solving and it will be the subject of this paper to survey some of these uses. Generally speaking, simulation projects are directed to one of the following objectives:

(1) System Design and Evaluation
(2) System Research and Planning
(3) Training—Appreciation, procedural or study display

Successful development in these areas will lead to the third major category of computer usage.

3. A CONTROLLING DEVICE

Currently, only exploratory pioneering work is being done in using the computer directly in management decision making. In this category of usage, we may visualize the computer as an "on-line" or "in-line" controller—operating on information received concerning sales, production, changes in environment, etc.—to make decisions on a day-to-day basis on personnel requirements, schedules for production, inventory pricing, etc.

To be effective, such usage involves the building of adequate decision-making criteria into a computer model of the company or of a component of it. The challenge, facing the would-be designer of what is referred to as a "truly integrated system," is to:

(1) Program a computer to become, as appropriately as possible, an on-line controller and,
(2) Make the computer an effective instrument for experimenting with, and evaluating the effectiveness of, proposed changes in policies, procedures, and plans.

While such automatic decision making (at the level of the firm) is still a long way off, significant strides are being made in some of the components and subcomponents of the business. Some of the simulation projects under the category of "system design and evaluation" performed in components of the business are clearly the forerunners of this new management control concept. One can gain much insight into the nature of future management control systems through study and examination of such projects.

II. PURPOSES IN THE USE OF SIMULATION

Simulation has long been used as an engineering method in the study and design of mechanisms and controlling apparatus.[2] Its extension to the managerial world has been for three major purposes.

[2] G. A. Hawkins, and L. M. K. Boelter, "The General Mode of Analysis in Engineering Education," *The Journal of Engineering Education*, 44:343–345, January, 1954, and H. H. Goode, "Simulation—Its Place in System Design," *Proceedings of the Institute of Radio Engineers*, 39:501–1506, 1951.

(1) To study complex operating plans and management controlling systems for the purpose of designing better plans and/or systems.

(2) To study and train people in the operation of complex tasks generally involving machines or instruments. Study of that old bugaboo—human interaction—is a major purpose in the use of simulation.

(3) To present proposed changes in such a way that enhances acceptance of change through better understanding of how a given system works or operates. In this way, the costly job of installation may be significantly reduced.

In this discussion, a control system will be defined as a machine and a set of procedures which directs, monitors, (i.e., controls) an operation. An operation is defined as an organization of men, machines, and information working toward a stated objective. There are several hierarchies of control systems and they may take different form in regard to their treatment of inputs and outputs. For example, a control system may have some automatic inputs and some manual inputs. It may have automatic outputs that provide closed loop inputs, or it may have manual outputs or displays for human monitoring.

III. TYPES OF SIMULATION

It is useful to further categorize simulation in regard to the presence or absence of competition in the basic model. A situation under study, involving the interactions of individuals, companies, or countries competing toward common, interrelated goals may be referred to as *Competitive Simulation.* Where the system, be it man or organization, is being studied in a changing environment or with changed system parameters—the term *System Simulation* may be applied. Examples of each may be found in the tabulation in the following section.

IV. THE USE OF SIMULATION IN TRAINING

Man has used simulating devices for many years to train people in the operation of new machines or equipment where training with the real equipment in real environment is either too costly or dangerous. In this discussion, which is primarily directed to use of simulation in management analysis, we shall pause only briefly to indicate that simulated situations often involving the use of a computer have been set up to provide training to individuals in the operation of management controlling systems or to provide simulated experience in decision making. A moment's consideration will bring out the importance of this point. We thrust the newly created manager into the role of operating a large organization without providing him with an opportunity to be trained in, or to experiment with, this most costly control apparatus. Simulation exercises have been designed to provide the executive trainee with an appropriate amount of synthetic experience in decision making.

These range from simple exercises, such as General Electric and Westinghouse have created to train distribution managers in the reorder rules in inventory control, to the top manage-

ment decision exercises of AMA,[3] IBM, and UCLA. The following is a partial listing of some simulation training exercises.

Organization	*Type of Simulation*	*Problem Area*	*Computer*
American Management Association and Booz, Allen and Hamilton	Competitive	Top Management Decision	Medium
General Electric Co.	System	Distribution Inventory	None
IBM Corp.	Competitive	Marketing Mgmt. Decision	Medium
RAND Corporation	System	Logistics System Training	Large
RAND Corporation		Monopologs	None
System Development Corporation	System	Air Defense System Training	SAGE
UCLA	Competitive	Top Management Decision	Medium
UCLA	System	Engineering Economy Classes	Small
Westinghouse Electric	System	Distribution Inventory	None

In passing, it is interesting to summarize the advantages claimed in using simulation in the management training area.

(1) Permits "cause" and "effect" to be felt.

(2) Aids student in evaluating available information.

(3) Familiarizes individual with data needed.

(4) Through "involvement," creates high motivation in student.

(5) Provides familiarity with Electronic Data Processing.

[3] R. Bellman, C. E. Clark, D. G. Malcolm, C. J. Craft and F. M. Ricciardi, "On the Construction of a Multi-Stage, Multi-Person Business Game," *Operations Research*, August, 1957, pp. 469–503.

V. INDUSTRIAL USES OF SIMULATION IN SYSTEM DESIGN AND EVALUATION

Simulation is useful in the study of a class of problems wherein the operating rules, policies, procedures, and other elements that control production, inventory, etc., are under question. One should like to employ practices that are the best and produce the smoothest, lowest cost operation. The number of variables involved, the uncertain nature of inputs, among other things,

make these problems, which are referred to generally as a system, difficult to analyze. A brief review of a few such projects will serve to illustrate the nature of the model and analysis.

1. MAN-MACHINE PRODUCTION OPERATION

Eastman Kodak has simulated their roll-film spooling operation, using Monte Carlo methods, on their 705 computer. Conditions change continuously in regard to this operation, both in the mix of size and run of individual products as well as in the maintenance requirements and equipment design. Simulation has been quite useful as the standard means for:

(1) Equipment redesign
(2) Organization of the operating crew procedures
(3) Maintenance and operating crew size determination

The model has permitted control methods to be developed so that better decisions can be made in the light of continuously changing conditions. This is but one of several simulations made by the Industrial Engineering Division.

The following is a listing of several simulation projects of the same general nature as those reported above.

SIMULATION IN SYSTEM DESIGN AND EVALUATION

Organization	*System Simulated*	*Computer*
Atlantic Refining	Inventory and Transportation of Casing	Medium
Eastman Kodak	Roll-Film Spooling Operation	Large
	In-Process Inventory	Large
	Elevator Systems, Plant Layout Study	Large
	Production Scheduling	Large
	Supply System	Large
General Electric	Job Shop Scheduling	Large
	Profit Planning Simulation	Medium
Humble Oil	Oil Tanker Scheduling	
Imperial Oil	Distribution System	Medium
Operations Research Office, The Johns Hopkins University	QM—Requirement Forecasting	Large
Port of New York Authority	Design of Bus Terminal	Medium
The RAND Corporation	Air Force Logistics System	Large
Thompson Products	Inventory Control	Large
United Air Lines	Customer Servicing	Medium
United Steel Cos. Ltd.	Steelworks Flow Problems	Special

VI. THE USE OF SIMULATION IN SYSTEM RESEARCH AND PLANNING

Projects in this area differ from the above largely in the fact that they are oriented more towards the future time frame and may consider the effect of equipment not yet designed. One of the distinct problems in evaluating the effect of introducing a technological change into an operating system is the subtle effect it may produce on operating costs. Lowering costs in one area often raises costs in other areas. The immense and time-consuming job of assessing these costs has often precluded the systematic study of alternatives. Computer Simulation can often be used as a means of determining *Total System Costs* for the change under study. Total system costs provide the most realistic basis for comparing alternative plans. An excellent example of such an application is:

1. NUCLEAR SHIP EVALUATION MODEL [4]

This problem, solved by Matson Navigation, compared ships of different sizes, speeds, and power plants, in addition to the proposed nuclear vessel whose design parameters were also to be determined. The problem was to determine where such a vessel might fit in the fleet of vessels serving a known cargo transportation pattern. A series of expected voyages representing more than a year of operations for the fleet was simulated. The cost and performance characteristics of each ship were entered into the computer and the fleet was operated according to predetermined assignment rules. The computer was programmed to keep track of costs, revenue, etc., and to permit evaluation to be made of different operational plans, different nuclear vessel design parameters and different mixes of types of ships.

Simulation projects of a similar nature are listed below.

SIMULATION IN SYSTEM RESEARCH AND PLANNING

Organization	*Problem Area*	*Computer*
Air Materiel Command	Aircraft Engine Management Model	None
Cal-Texas Oil Company	Group Operations Decision Model	None
Operations Research Office, The Johns Hopkins University	Army Battalion Tactical Maintenance	Large

VII. RESEARCH AT UNIVERSITIES AND GOVERNMENT AGENCIES

Any comprehensive discussion of research in this area would be prohibitive

[4] F. B. Graham, "Determining the Comparative Economics of Nuclear Propulsion," *Recent Research in Maritime Transportation*, Publication 592, National Academy of Sciences—National Research Council, Washington, 1958.

in length. Suffice it to say that the simulation approach is proving quite useful to the basic researcher in more adequately describing problems of a dynamic nature. The simulation approach has proved a useful tool in conserving our scarce mathematical talents. Often laying out the problem in this manner gives sufficient structure to the problem to challenge the talents of creative mathematicians who otherwise might not be sufficiently convinced that there was a problem requiring their talents.

The following listing is offered as being indicative of the types of simulation projects underway.

Organization	*Problem Area*
Canadian Defence Research Board	Operations in Mining Cycle
George Washington University	Tanker Design
MCTC	Port Operations
M.I.T.	Inventory Control Model Vehicular Traffic
Stanford	Evaluation of Quality Control Plans
Tufts	Production Scheduling
U.C.L.A.	Cargo Handling Containerization Warehouse Layout

VIII. PROBLEMS IN USE OF SIMULATION

Perhaps the most dramatic aspects of system simulation lie in its ability to reproduce the workings of large-scale systems. While such ambitious programs are of great interest, it should be borne in mind that their success is dependent on having performed many smaller simulations for experience, upon adequate input data, and upon consideration of the many mathematical and statistical problems that are involved in the construction of a useful model. Those considering broad full-scale models as panaceas in the management analysis and decision-making areas are well advised to engage in smaller projects at the outset. This will prevent certain disillusionment and make for a more meaningful research program.

Following this vein, a listing of some of the more common problems to be encountered in the use of simulation may be useful.

(1) Broad problems generally require the use of a computer. This can be costly both from a programming as well as an operating point of view. For example, one of the large-scale simulations, referred to here, took over two years to become operational.

(2) Development of considerable new data is generally necessary. The input distributions required in simulations cannot always be derived from existing data sources and surveys. Quite often experiments must be designed to obtain the necessary data. While costly, this data collection usually develops other information and insights of great value.

(3) Very large problems are often unwieldy and hard to program. The interrelationship of factors in the model adds considerably to the complexity.

(4) In large problems, the task of exploring all the possibilities of parameter changes creates a volume of calculations that may swamp the analyst. Consideration should be given to analysis before designing the simulation.

(5) Comparison of simulation runs, as well as their length, pose statistical problems requiring the presence of an experienced mathematical statistician.

(6) The effect of the accuracy of input data should be explored. Analysis of outputs of a simulation based on input data of unknown or questionable inputs is difficult.

(7) And finally, in their enthusiasm for this method, many analysts have discovered that simpler methods of analysis may exist. It is well to search for such methods early in the task of problem solving.

IX. THE FUTURE OUTLOOK

It seems clear that modern management is driving continuously toward a better understanding of the "process" for which it is responsible. In this search for more precise definition concerning what is controllable and what is uncontrollable, it is quite evident that simulation modeling will be a necessary research approach. As the answers unfold we shall gradually see develop a new management control concept based on precepts discovered in the operation of the research models.

In the future it does not seem at all unlikely that management will have a computer model of its business, rich enough in detail and comprehensive enough in scope to permit experimentation with suggested policy change. Further, the model may well be able to administer policy more adequately and consistently than the human administrator. The decision maker will then be freed for the more important task, that of understanding the limitations of the model and of searching imaginatively for beneficial innovation.

✦✦✦✦✦✦✦✦✦✦✦✦✦ CASES IN SIMULATION: *a research aid as a management "demonstration piece"*

PATRICK J. ROBINSON

The Operations Research method called Systems Simulation is a relatively recent development and may not appear in dictionaries yet. Perhaps the following discussion will help you appreciate how it may be of value to you in your business.

In many cases we would like to know in advance something about the probable behavior of a business operation or competitive system before we actu-

From an address at the Illinois Institute of Technology, June 1957.

ally start it. We cannot usually arrange to have an experiment conducted using an entire business operation, simply because it would take too long, be too costly or be contrary to Company policy, and so forth. About the only way that we can hope to predict the consequences of our actions before committing ourselves, is through intuitive business judgment, scientific study, or possibly sheer speculation. Probably the best approach is to apply a useful scientific method to support experienced management judgment. This is where we come to the possibility of simulation—particularly if we are dealing with a changing or dynamic situation with some complexity.

A general illustration is probably as good a way as any to get this point across. We may try something out on a small scale as an experiment or "straw in the wind" or do it in a way we often refer to as a "dry run." This type of imitation of reality, in an attempt to see what might happen under conditions of real operation by doing a test on paper or in some other artificially limited fashion, is what can be referred to as Systems Simulation. We can attempt to do a series of experiments on a restricted basis, perhaps in a laboratory testing a physical model of the system. In Management Science we would usually use a manual or electronic computer to test a mathematical model which can give us the feel and simulated results of real business activities without anything like the time, expense and, more importantly, risks involved in experimenting with the actual business.

Simulation is something which stems from Scientific Method which is basic to all true research. This type of activity is merely the outcome of trying to reason in a strictly logical fashion from initial hypotheses through various carefully designed experiments. What we try to do in simulating for research purposes is to develop first a crude idealized system or model, and then through successive approximations and refinements we strive for a usefully close reflection of actuality. However, whether it is a model of a new aircraft, or a mathematical model of a business or military operation, the model is put through its paces by being subjected to test conditions which closely resemble or simulate conditions which are anticipated for actual operation. By observing carefully the behavior of the model, whether it be physical or mathematical, we can gather some real insight into the likely behavior of the whole system under actual operating conditions. Naturally, if our simulation lacks certain basic elements of realism, or has distorted key relationships, we cannot hope to have a simulation worth using. In fact, herein lies one of the principal dangers to be avoided. If a thorough study cannot be made, it may be better to rely on judgment alone than to risk being misled by inadequate research.

You know that Link Trainers or other flight simulators are not primarily test models in the sense of wind tunnel models, for the design of prototypes. Nevertheless, these devices and their users constitute another true form of simulation for other vital purposes. This type of simulation is for the training of men who must learn to make ef-

fective decisions and take prompt and proper action. It is far faster, safer and more productive of thorough training to simulate reality as closely as possible under controlled test conditions before undertaking the actual task being learned.

You have probably read in *Nation's Business* or elsewhere of the American Management Association's Competitive Business War Games. They are another outstanding form of simulation for training purposes. In such games the training is not in the use of equipment or machines, but rather in gaining a grasp and appreciation of the interactions of competitive business strategies and the value of effective planning and profitable tactics from a policy making top management viewpoint.

Eventually, such management decision games may prove useful beyond the vital training area by helping solve actual problems and even help develop winning competitive strategies. However, reliable help beyond training may be some years ahead and must be approached with caution.

Now we may touch on the dual nature of the title of this paper. We are speaking of a research aid that helps us gain an insight into business operations and, at the same time, can become a demonstration piece, or display of probable results, for discussion or training purposes. Often, responsible business administrators must be shown the reasoning and basis for observations and analysis if they are to be confident of the research and prepared to take action, whether it be training or planning strategies using Systems Simulation. It is management's prerogative and, in fact, often the saving grace of good sound business judgment, that "why" questions are asked concerning observations and recommendations made by management's advisors. This gives rise to a need for mutual understanding and good communication featuring an ability to illustrate ideas clearly and appealingly.

In research work simulation is a valuable tool of great value to the researcher. Equally important, it can also be a useful means of demonstrating to business administrators what likely behavior can be expected under certain practical conditions as a basis for more effective decision-making.

You will be interested in what types of simulation applications can be made that would be of direct interest in your own businesses and it's probably safe to say that there are hardly any businesses in which applications are not conceivable. However, as with many other techniques, there is a need to be prudent as to where a technique such as this can be helpful or where it is merely apt to be applied as a fad. There is some danger of new tools being taken around by keen young men searching for matching problems; like accidents looking for a place to happen! Clearly, this is an "upside down" approach.

Suffice it to say that wherever an operation is sufficiently complex and where the risks or cost of trial and error methods on a large scale are such that there is a great incentive to do a careful prestudy of alternatives covering dynamic conditions, which may be hard to predict, there are likely grounds for applying simulation as a profitable

method of Business Research as an aid to planning and training.

You may now find it helpful to consider three specific applications of simulation to problem solving and forecasting of the benefits or penalties involved in three related phases of a business operation. These cases may be highlighted best through the use of slides [several of which are reproduced in the original paper]. They concern the ordering, storing and distribution of products starting at a manufacturing level and going forward to the point of ultimate sale.

Some years ago the Company had a problem which appeared to stem from the need for enlarged and modernized field warehouses all across Canada. Since there were many hundred field warehouses involved, the cost of an extensive expansion program would have been very high indeed. The problem was to determine what was required and how best to go about meeting the difficulties arising from severe overcrowding at many field plants. Preliminary examination revealed that in most instances the field warehouses were overstocked to the point of inefficiency. In addition, the garage facilities were generally taken over as additional storage space and in many instances the yards around the plants were crowded with barrels which were left out-of-doors. Many problems arose as a result of these conditions. For example, in some cases there was more than a year's supply of an item on hand at one location. In other cases, runouts were occurring and emergency arrangements were necessary to fill customer requirements. Costly transfers or very rushed orders from the central supply point in Eastern Canada were frequent. The customer was being well served—but at an excessively high cost to Imperial Oil Limited.

These field warehouses stock all types of packaged petroleum products and also tires, batteries and accessories. We need not go into great detail about the eight or nine man-years' of work that has gone into our research activities on inventory control and ordering systems. Basically, the need of a major central warehousing function in the case of packaged petroleum products was established and a suitable Automatic Inventory Control System devised. In addition a companion inventory control system was developed for use in field warehouses. These systems were designed to dovetail so that the system applied at the central point serviced its field plant customers in much the same way that the individual field plant warehouses handled their service stations, farm trade agents and other customers. The systems were intended to provide for adequate insurance against runout conditions, while ensuring low cost operation and efficient activities at the least capital outlay.

Since much of the cushion stocks had previously been carried in the field plants, there were substantial reductions potentially available through centralizing reserves and thus reducing the insurance stocks carried at the isolated plants. In addition, high speed filling facilities permitted much of the central stocks to be concentrated in bulk in semi-finished form. Furthermore, complete carload, rather than less than car or van load shipments were facilitated.

Essentially, a relatively uniform manufacturing operation was being protected on the one hand, and a fluctuating customer demand was being protected on the other. In between, complex supply networks were carrying the necessary products across the length and breadth of Canada.

The same type of operation existed for the stocks of tires and related items but in this instance the products were being purchased from a supplier who was doing the central warehousing. This changed the problem somewhat and a different solution was attained. We will discuss this later.

Since it was determined that it was essential to construct a central facility, and establish a reliable central supply system before initiating similar systems in the field plants, a number of man-years of research effort went into the study and installation of these facilities and systems to aid the operating personnel concerned. The problem of demonstrating to management at the outset that a huge central warehouse was indeed a desirable thing, was handled using several means. One aid was a device which might be referred to as a glorified pinball machine and which might be illustrated by the viewgraph that is a take-off on this device. This simple electric and mechanical model illustrated the basic features of the inventory control system. This was definitely a demonstration device and had not been at any point used for research purposes. However, in order to provide further information on refinements of the integrated system at the central facility, and to ensure that any major bottlenecks would be spotted in the preliminary stages of planning rather than in actual operation, some means of actively demonstrating the workings of the system under various conditions had to be devised. This was particularly essential and difficult since no central warehouse facility had existed previously and, as a result, there was nothing suitable to use for comparative purposes. However, after some thought we decided to program a large-scale electronic computer at the University of Toronto on which to base our simulation of the entire central warehousing system.

First, we worked with management in the Manufacturing Operation to develop a detailed flow diagram which represented the system and the decisions that would have to be made. Then we worked with the Computation Center personnel to devise a program of instruction for the computer, such that we could provide it with all the basic information, control points, and other pertinent data including various practical constraints. With this information in the machine, we could give the computer an initial set of inventory levels for the many hundreds of items being stocked, and provide it with a detailed daily recapitulation of orders from field plants that had taken place in actual operation in previous time periods. Data for a number of critical months were particularly useful in testing the system for possible shortages. The computer was programmed to print periodic stock reports to indicate the inventory levels, and in addition to provide information on all unusual conditions such as shortages or waiting lines for facilities. These were then an-

alyzed so that refinements could be introduced into the actual operation.

One interesting sidelight at this stage was that this was in 1953 and was our first large-scale electronic computer program and it taught us a few lessons. We experienced considerable difficulty in developing an adequately detailed program and even more difficulty in obtaining all of the necessary data. Undoubtedly many of you gentlemen have experienced this type of overwhelming experience when dealing with computers for the first time. The need for clear instructions covering every possible eventuality cannot be over-emphasized. Finally, we were able to run the program and obtain the results somewhat before the actual operation went into effect. Fortunately, management had decided to go ahead on the basis of preliminary indications and the simple "pinball simulation" and various paper analyses.

Some time later, after the construction of the central warehouse, and during installation of the automatic inventory control system, new manufacturing and central storage facilities were built in Western Canada and parallel assistance was provided in installing suitable control systems. At this juncture we were in the position of having to demonstrate to Marketing operating personnel the value and anticipated results from installing the dovetailing inventory control systems in field warehouses. One of the ways to show what to expect, since field warehouses had been in operation for years, was to simulate a field warehouse operation using actual data. As these operations were on a much smaller scale we handled the simulation through manual calculations. The interesting feature about simulating these operations on paper and using actual recapitulations of data to test the system was that in the initial test we were provided with information from plants which, under tests, suggested that we could not anticipate appreciable savings. On examining this closely and reviewing it with the operating people concerned, we were informed, to our relief, that they had selected from the hundreds of plants available the data covering the two best operations at their disposal. Apparently their theory had been that if we could in any way approximate by a system, what they knew to be a good operation by a long experienced operator, we would have demonstrated the system's stability and reliability. We were very close to showing similar stock levels in all categories, and in fact, were able to indicate a lower number of runout conditions than had actually occurred. This is illustrated also in the various Vugraphs that apply. Subsequent simulations of other plants showed that substantially greater savings could be anticipated, and these savings were a reflection of putting in a good system which could be handled readily by a new man in contrast to the actual operations where not every plant had men with anything like the experience that the first two under tests had had.

The preceding two illustrations are based on the handling of packaged petroleum products. The related system which we will now consider deals with non-petroleum products in the tires, batteries and accessories lines. These

commodities occupy the same warehouses and, in fact, have preferred space because of their perishable nature. Every effort to reduce the quantities of these tied up, and the consequent over-crowding suggested the need for studies along the lines of our work in the other inventory control areas. The results of a fair amount of investigation emphasized the need for a basically new ordering system rather than a central warehouse, since the suppliers were providing reasonable service in this respect.

First, a flow plan of the entire system was constructed. This approach may be quite familiar to many of you since it is quite similar to the sort of thing that can be drawn for materials flow, but here the flow indicates decisions. These flow plans have very interesting properties since they indicate not only direct information flow, but also recirculation or "feedback" of information. This might be compared with the situation that you are confronted with at home with a fuel oil burner in the basement and a thermostat indicator which is set to maintain a certain temperature in a room. If you set the thermostat to 70° this sensing device merely instructs the burner whether to go on or off to hold 70° (plus or minus some designed tolerance) and the feedback of information to the burner is what controls its activity. If the temperature goes up too high information is fed back to the burner to stop, and contrariwise the burner is turned on when the temperature falls too low. This type of activity and much more complex servo-mechanisms such as automatic pilots in aircraft can be analyzed mathematically by applying what is known as servo-mechanism theory. This is a well-established body of techniques that was built up primarily to serve design engineers of complex servo-mechanism devices such as refinery controls, heating systems and innumerable other devices in common use today.

Several interesting contrasts were soon discovered in this product ordering study. It seemed preferable to order on a regular time basis, that is, once a month or some other fixed date, rather than at predetermined stock levels as with the packaged petroleum products where the interval between orders was not fixed but the order quantities were. A certain amount of rudimentary forecasting was possible in the case of seasonal customer demand and this could be usefully built into a system of checks and balances which could permit the periodic revision of orders to correct for over or under estimating.

We need not go into details on computations and methods employed other than to say that several operations research techniques, including Monte Carlo methods and various aspects of probability theory, were employed. We can say that we were again confronted with the problem of demonstrating to the operating management concerned, the merits of any system of this sort. However, before demonstrating it to their satisfaction, we had to develop a series of research efforts which were intended to provide us with a better understanding of the systems under study. These analyses took the form of constructing mathematical models based on the servo-mechanisms de-

scribed. Through a series of successive approximations, each of which was intended to be closer to reality, we were able to devise a system which was useful for predicting the probable consequences of alternate courses of action under different conditions of demand and uncertainty. Incidentally, it is well to remark here that in the tests of the other inventory control systems not only actual data were employed, but data reflecting predictably higher sales and other occurrences were used to test the model for sensitivity to deviations beyond the limits for which the systems were designed. It is very important to make sure that every possible contingency has been included in a reasonable fashion and that successive testing of the model and the system will reveal any weaknesses and permit a rather close reflection of actual operating conditions. Once a reasonably reliable model has been developed then one is in the position of having a basis for sequential testing of different conditions.

This type of approach has a great deal to be said for it. Furthermore, the complexities are often beyond the power of straight analytic means and can best be handled through repeated tests which reproduce five, ten or even one hundred years of experience in a short space of time. In this particular case, 88 months of historical data were used to run through the model in simulation. During these 88 months various information was accumulated indicating the stock turnover, the number of runouts and various considerations as to inventory levels and general performance. It was interesting to see that the designed tolerances on inventory coverage were satisfied and in addition, significant stock reductions were indicated. It is completely out of the question to be 100% protected against stock shortages. If you think about this you will appreciate that it is a virtually impossible and enormously expensive goal to aim for. Somehow, one must determine what the cost of an unsatisfied customer may be. Questions of "Do you lose the profits from the one sale?" or "Do you lose all of this man's future trade?" or "Do you lose not only this type of business from this source, but also related business?," tie-in-sales?—all these things have to be considered in evaluating what degree of protection you can reasonably afford to provide. Here again in the simulation tests it was interesting to see that a reduced number of shortages were indicated despite these lower levels of stock, owing to the improved systematic controls and checks and balances built into the system.

We were particularly gratified in the installation of this system since not only were the simulation methods very successful in conducting the research and designing a really effective system, but also in demonstrating to management the results to be expected. Possibly of the most importance was the use of these methods subsequently in selling the men in the plants and all through the operation. In the final analyses, these are the ones who really have to be convinced before any system can be expected to receive a fair trial, much less a successful installation.

In summing up, we might draw a few conclusions concerning simulation

methods. It is clear that in many cases where complex relationships, both of predictable and random natures, occur, it is easier to set up and run through a simulated situation than it is to develop and use a mathematical model representing the entire process under study. In many cases an activity can be affected by numerous random influences. The probabilities involved for each type of influence can be separately examined. However, the calculation of the probability of the combined sequence of activities spilling over into each other and interacting, in what is sometimes referred to as a cascading effect, leads into very deep mathematical waters in the field of Stochastic Processes. This is simply a sophisticated technical phrase or piece of jargon covering involved problems of uncertainty.

To avoid an impossible or unprofitable attempt to solve a complex operation using equations to seek so-called optimal answers, we turn to a simulation of a system such that we may repeatedly experiment and obtain statistically reliable empirical results. Sometimes we may use simulation or Monte Carlo Techniques (this name is aptly suggestive) to firm up or verify complex theoretical analyses. A good synthesis of an operation might be thought of as a form of mathematical "sausage machine," grinding out the results for as long as we care to feed material into it. It is not simply a "better mousetrap" providing a "one-shot" proposition. It does not simply provide one answer. Rather it sets up a general framework within which we can, in a series of tests, run through months or years of experience to see what the likely outcome will be.

As we have seen and you may appreciate intuitively, simulation can be of considerable value for training purposes. It can also be useful in basic experimentation and in the evaluation of various problems. In addition, we have seen that simulation techniques can tackle problems that are too cumbersome to be handled analytically. It can handle major systems or deal with rather small ones. The synthesis of a system is not dependent on how many details are put into it, but how many important factors are included. As long as no significant omissions exist, and as long as there is some empirical understanding of the various actions and interactions, a satisfactory model can probably be built for running simulation tests.

Where no analytic solution is available, the search for an ever-improving answer through the sequential solution of alternate trials, until finally running and rerunning additional cases doesn't produce any material improvement, brings us near to what we can, with confidence, rely on as being something approximating an optimal solution. The input data to a simulation may be real or generated data. There may be a number of components in this simulation. We don't necessarily have to rely only on computers or desk calculators; we can, in the case of, say, simulation of a weapons system or some other complex arrangement, build people into the system who will respond to various stimuli and will be part of the overall test. There are all manner of computers which may be employed in simulation.

Digital computers with suitable programs are the types of equipment which generally will be applied, but all manner of special purpose analogue or mixed equipment is feasible.

There are basically different types of simulation programs. The type which preprograms everything that is known about a situation, and fills in the answers to all the probable conditions which may arise, is very effective, although not as sophisticated as the type of program which permits the machine to do some form of elementary learning, such as a machine which learns the rules and plays a simple game of chess, or a simulation in which a certain amount of trial and error generates a body of experience which results in an improved play the next time a similar situation arises. The only basic difference in what we get from a simulation from what we get from an actual operation is that from a simulation the output is in the form of data which must be interpreted, and in an actual operation we get operating reports and frequently it is too late to do anything about recouping losses—we simply have to try to avoid similar losses again, or if we make a substantial profit somewhere we must try to determine what we did to get it, so that we may do it again.

Summing up, you may agree that through simulating reality on a modest scale we can do some research that might not be feasible otherwise; while at the same time we may provide management with a dramatic demonstration piece to help appraise situations and make profitable decisions.

part ✦ 4

Some Applications of Operations Research

Some examples of operations research at work have already been mentioned and described in articles in the previous section of this book. Since the emphasis in that section has been on the nature of the techniques, the examples of their application have been rather sketchy. The full flavor of the OR approach to a problem may not have been transmitted to the reader. To assure, therefore, a more certain and keener appreciation of OR in action, this concluding section of the book has been constructed of two kinds of articles: (1) fuller descriptions of some actual OR studies; and (2) suggestions of operations analysts pertaining to improvements in the quality of particularly difficult business decisions through the application of their special techniques.

Today, the literature of operations research abounds with articles of the types mentioned. Unfortunately, however, the preponderant majority of these articles require for their understanding greater mastery of the techniques of OR than could have been obtained from this book. In a sense, therefore, the articles which follow are not truly representative of the OR work they are intended to represent. Nevertheless, these articles cannot but provide the reader with a deeper appreciation and understanding of the OR approach to business problems and that, after all, is the aim here.

In selecting the articles for this section an effort was made to provide illustrations in each of the major functional areas of business: production, marketing, and finance. Thus, the first three articles—two of which were written by the Research Director of the Operations Research Group of Arthur D. Little, Inc., and the third by the Director of the OR Group at the Case Institute of Technology—describe work in production planning and inventory control. These same two men are also the authors of the next two articles but this time the work described is in

the marketing area. In addition, there is a third article dealing with an important problem of marketing management. This article was authored by the President of Harvey Shycon, Inc., a firm of marketing consultants, and by a member of the School of Industrial Management of the Massachusetts Institute of Technology.

Finally, the last three articles describe approaches to problems in finance. The first is in the area of accounting and was authored by a prominent member of the U.S. Internal Revenue Service. The second, on the other hand, sheds some new light on one of the thorniest problems in financial management; capital budgeting. Its author is Consulting Economist for the General Economics Department of Standard Oil Company of New Jersey. The final article in this section—written by the principal-in-charge of the Management Services Central Staff of Touche, Ross, Bailey and Smart, a leading accounting firm—offers an approach to one of the most critical as well as most difficult problems executives must face, the problem of mergers or acquisitions.

A Operations Research in Production Management

GUIDES TO INVENTORY POLICY: *functions and lot sizes*

JOHN F. MAGEE

"Why are we always out of stock?" So goes the complaint of great numbers of businessmen faced with the dilemmas and frustrations of attempting simultaneously to maintain stable production operations, provide customers with adequate service, and keep investment in stocks and equipment at reasonable levels.

But this is only one of the characteristic problems business managers face in dealing with production planning, scheduling, keeping inventories in hand, and expediting. Other questions —just as perplexing and baffling when managers approach them on the basis of intuition and pencil work alone— are: How often should we reorder, or how should be adjust production, when sales are uncertain? What capacity levels should we set for job-shop operations? How do we plan production and procurement for seasonal sales? And so on, and so on.

In this series of articles, I will describe some of the technical developments which aim at giving the business manager better control over inventory and scheduling policy. While these techniques sometimes employ concepts

Harvard Business Review, *January–February 1956, 34:1, 49–60.*

and language foreign to the line executive, they are far from being either academic exercises or mere clerical devices. They are designed to help the business manager make better policy decisions and get his people to follow policy more closely.

As such, these techniques are worth some time and thought, commensurate with the central importance of production planning and inventory policy in business operations. Indeed, many companies have found that analysis of the functions of inventories, measurement of the proper level of stocks, and development of inventory and production control systems based on the sorts of techniques described in this and following sections can be very profitable. For example:

Johnson & Johnson has used these techniques for studying inventory requirements for products with seasonally changing demand, and also to set economical inventory goals balancing investment requirements against additional training and overtime costs.

The American Thread Company, as a supplier to the fashion goods industry, plagued with large in-process inventories, day-to-day imbalances among production departments, labor turnover, and customer service difficulties, found these methods the key to improved scheduling and control procedures. Now these improved procedures help keep an inventory of tens of thousands of items in balance and smooth out production operations even in the face of demand showing extremely erratic fluctuations due to fashion changes.

The Lamp Division of the General Electric Company has reported using these methods to survey its finished inventory functions and stock requirements in view of operating conditions and costs. This survey indicated how an improved warehouse reorder system would yield inventory cuts at both factories and warehouses, and pointed to the reorder system characteristics that were needed; it led to the installation of a new reorder and stock control system offering substantial opportunities for stock reduction. The analytic approach can also be used to show clearly what the cost in inventory investment and schedule changes is to achieve a given level of customer service.

An industrial equipment manufacturer used these methods to investigate inventory and scheduling practices and to clear up policy ambiguities in this area, as a prelude to installing an electronic computer system to handle inventory control, scheduling, and purchase requisitions. In general, the analytic approach has proved a valuable help in bringing disagreements over inventory policy into the open, helping each side to recognize its own and the others' hidden assumptions, and to reach a common agreement more quickly.

The Procter & Gamble Company recently described how analysis of its factory inventory functions and requirements, using these methods, has pointed out means for improved scheduling and more efficient use of finished stock. The analysis indicated how the company could take advantage of certain particular characteristics of its factories to cut stocks needed to meet sales fluctuations while still maintaining its long-standing policy of guaranteed annual employment.

These are only a few instances of applications. Numerous others could be drawn from the experience of companies ranging from moderate to large size, selling consumer goods or industrial products, with thousands of items or only a few, and distribution in highly

stable, predictable markets or in erratically changing and unpredictable circumstances.

In the present article major attention will be devoted to (a) the conceptual framework of the analytic approach, including the definition of inventory function and the measurement of operational costs; and (b) the problem of optimum lot size, with a detailed case illustration showing how the techniques are applied.

This case reveals that the appropriate order quantity and the average inventory maintained do not vary directly with sales, and that a good answer to the lot size question can be obtained with fairly crude cost data, provided that a sound analytical approach is used. The case also shows that the businessman does not need calculus to solve many inventory problems (although use has to be made of it when certain complications arise).

INVENTORY PROBLEMS

The question before management is: How big should inventories be? The answer to this is obvious—they should be just big enough. But what is big enough?

This question is made more difficult by the fact that generally each individual within a management group tends to answer the question from his own point of view. He fails to recognize costs outside his usual framework. He tends to think of inventories in isolation from other operations. The sales manager commonly says that the company must never make a customer wait; the production manager says there must be long manufacturing runs for lower costs and steady employment; the treasurer says that large inventories are draining off cash which could be used to make a profit.

Such a situation occurs all the time. The task of all production planning, scheduling, or control functions, in fact, is typically to balance conflicting objectives such as those of minimum purchase or production cost, minimum inventory investment, minimum storage and distribution cost, and maximum service to customers.

PRODUCTION VS. TIME

Often businessmen blame their inventory and scheduling difficulties on small orders and product diversity: "You can't keep track of 100,000 items. Forecasts mean nothing. We're just a job shop." Many businessmen seem to feel that their problems in this respect are unusual, whereas actually the problems faced by a moderate-size manufacturer with a widely diversified product line are almost typical of business today.

The fact is, simply, that under present methods of organization the costs of paper work, set-up, and control, in view of the diversity of products sold, represent an extremely heavy drain on many a company's profit and a severe cost to its customers. The superficial variety of output has often blinded management to the opportunities for more systematic production flow and for the elimination of many of the curses of job-shop operation by better organization and planning.

The problem of planning and scheduling production or inventories per-

vades all operations concerned with the matter of production versus time—i.e., the interaction between production, distribution, and the location and size of physical stocks. It occurs at almost every step in the production process: purchasing, production of in-process materials, finished production, distribution of finished product, and service to customers. In multiplant operations, the problem becomes compounded because decisions must be made with reference to the amount of each item to be produced in each factory; management must also specify how the warehouses should be served by the plants.

ACTION VS. ANALYSIS

The questions businessmen raise in connection with management and control of inventories are basically aimed at action, not at arriving at answers. The questions are stated, unsurprisingly, in the characteristic terms of decisions to be made: "Where shall we maintain how much stock?" "Who will be responsible for it?" "What shall we do to control balances or set proper schedules?" A manager necessarily thinks of problems in production planning in terms of centers of responsibility.

However, action questions are not enough by themselves. In order to get at the answers to these questions as a basis for taking action, it is necessary to back off and ask some rather different kinds of questions: "Why do we have inventories?" "What affects the inventory balances we maintain?" "How do these effects take place?" From these questions, a picture of the inventory problem can be built up which shows the influence on inventories and costs of the various alternative decisions which the management may ultimately want to consider.

This type of analytic or functional question has been answered intuitively by businessmen with considerable success in the past. Consequently, most of the effect toward improved inventory management has been spent in other directions; it has been aimed at better means for recording, filing, or displaying information and at better ways of doing the necessary clerical work. This is all to the good, for efficient data-handling helps. However, it does not lessen the need for a more systematic approach to inventory problems that can take the place of, or at least supplement, intuition.

As business has grown, it has become more complex, and as business executives have become more and more specialized in their jobs or farther removed from direct operations, the task of achieving an economical balance intuitively has become increasingly difficult. That is why more businessmen are finding the concepts and mathematics of the growing field of inventory theory to be of direct practical help.

One of the principal difficulties in the intuitive approach is that the types and definitions of cost which influence appropriate inventory policy are not those characteristically found on the books of a company. Many costs such as setup or purchasing costs are hidden in the accounting records. Others such as inventory capital costs may never appear at all. Each cost may be clear to the operating head primarily

responsible for its control; since it is a "hidden" cost, however, its importance may not be clear at all to other operating executives concerned. The resulting confusion may make it difficult to arrive at anything like a consistent policy.

In the last five years in particular, operations research teams have succeeded in using techniques of research scientists to develop a practical analytic approach to inventory questions, despite growing business size, complexity, and division of management responsibility.

INVENTORY FUNCTIONS

To understand the principles of the analytic approach, we must have some idea of the basic functions of inventories.

Fundamentally, inventories serve to uncouple successive operations in the process of making a product and getting it to consumers. For example, inventories make it possible to process a product at a distance from customers or from raw material supplies, or to do two operations at a distance from one another (perhaps only across the plant). Inventories make it unnecessary to gear production directly to consumption or, alternatively, to force consumption to adapt to the necessities of production. In these and similar ways, inventories free one stage in the production-distribution process from the next, permitting each to operate more economically.

The essential question is: At what point does the uncoupling function of inventory stop earning enough advantage to justify the investment required? To arrive at a satisfactory answer we must first distinguish between (a) inventories necessary because it takes time to complete an operation and to move the product from one stage to another; and (b) inventories employed for organizational reasons, i.e., to let one unit schedule its operations more or less independently of another.

MOVEMENT INVENTORIES

Inventory balances needed because of the time required to move stocks from one place to another are often not recognized, or are confused with inventories resulting from other needs—e.g., economical shipping quantities (to be discussed in a later section).

The average amount of movement inventory can be determined from the mathematical expression $I = S \times T$ in which S represents the average sales rate, T the transit time from one stage to the next, and I the movement inventory needed. For example, if it takes two weeks to move materials from the plant to a warehouse, and the warehouse sells 100 units per week, the average inventory in movement is 100 units per week times 2 weeks, or 200 units. From a different point of view, when a unit is manufactured and ready for use at the plant, it must sit idle for two weeks while being moved to the next station (the warehouse); so, on the average, stocks equal to two weeks' sales will be in movement.

Movement inventories are usually thought of in connection with movement between distant points—plant to warehouse. However, any plant may contain substantial stocks in movement

from one operation to another—for example, the product moving along an assembly line. Movement stock is one component of the "float" or in-process inventory in a manufacturing operation.

The amount of movement stock changes only when sales or the time in transit is changed. Time in transit is largely a result of method of transportation, although improvements in loading or dispatching practices may cut transit time by eliminating unnecessary delays. Other somewhat more subtle influences of time in transit on total inventories will be described in connection with safety stocks.

ORGANIZATION INVENTORIES

Management's most difficult problems are with the inventories that "buy" organization in the sense that the more of them management carries between stages in the manufacturing-distribution process, the less coordination is required to keep the process running smoothly. Contrariwise, if inventories are already being used efficiently, they can be cut only at the expense of greater organization effort—e.g., greater scheduling effort to keep successive stages in balance, and greater expediting effort to work out of the difficulties which unforeseen disruptions at one point or another may cause in the whole process.

Despite superficial differences among businesses in the nature and characteristics of the organization inventory they maintain, the following three functions are basic:

(1) *Lot size inventories* are probably the most common in business. They are maintained wherever the user makes or purchases material in larger lots than are needed for his immediate purposes. For example, it is common practice to buy raw materials in relatively large quantities in order to obtain quantity price discounts, keep shipping costs in balance, and hold down clerical costs connected with making out requisitions, checking receipts, and handling accounts payable. Similar reasons lead to long production runs on equipment calling for expensive setup, or to sizable replenishment orders placed on factories by field warehouses.

(2) *Fluctuation stocks*, also very common in business, are held to cushion the shocks arising basically from unpredictable fluctuations in consumer demand. For example, warehouses and retail outlets maintain stocks to be able to supply consumers on demand, even when the rate of consumer demand may show quite irregular and unpredictable fluctuations. In turn, factories maintain stocks to be in a position to replenish retail and field warehouse stocks in line with customer demands.

Short-term fluctuations in the mix of orders on a plant often make it necessary to carry stocks of parts of subassemblies, in order to give assembly operations flexibility in meeting orders as they arise while freeing earlier operations (e.g., machining) from the need to make momentary adjustments in schedules to meet assembly requirements. Fluctuation stocks may also be carried in semifinished form in order to balance out the load among manufacturing departments when orders received during the current day, week, or month may put a load on individual departments which is out of balance with long-run requirements.

In most cases, anticipating all fluctuations is uneconomical, if not impossible. But a business cannot get along without some fluctuation stocks unless it is willing and able always to make its customers wait until the material needed can be purchased conveniently or until their orders can be scheduled into production conveniently.

Fluctuation stocks are part of the price we pay for our general business philosophy of serving the consumers' wants (and whims!) rather than having them take what they can get. The queues before Russian retail stores illustrate a different point of view.

(3) *Anticipation stocks* are needed where goods or materials are consumed on a predictable but changing pattern through the year, and where it is desirable to absorb some of these changes by building and depleting inventories rather than by changing production rates with attendant fluctuations in employment and additional capital capacity requirements. For example, inventories may be built up in anticipation of a special sale or to fill needs during a plant shutdown.

The need for seasonal stocks may also arise where materials (e.g., agricultural products) are *produced* at seasonally fluctuating rates but where consumption is reasonably uniform; here the problems connected with producing and storing tomato catsup are a prime example.[1]

STRIKING A BALANCE

The joker is that the gains which these organization inventories achieve in the way of less need for coordination and planning, less clerical effort to handle orders, and greater economies in manufacturing and shipping are not in direct proportion to the size of inventory. Even if the additional stocks are kept well balanced and properly located, the gains become smaller, while at the same time the warehouse, obsolescence, and capital costs associated with maintaining inventories rise in proportion to, or perhaps even at a faster rate than, the inventories themselves. To illustrate:

Suppose a plant needs 2,000 units of a specially machined part in a year. If these are made in runs of 100 units each, then 20 runs with attendant setup costs will be required each year.

If the production quantity were increased from 100 to 200 units, only 10 runs would be required—a 50% reduction in setup costs, but a 100% increase in the size of a run and in the resulting inventory balance carried.

If the runs were further increased in length to 400 units each, only 5 production runs during the year would be required—only 25% more reduction in setup costs, but 200% more increase in run length and inventory balances.

The basic problem of inventory policy connected with the three types of inventories which "buy" organization is to strike a balance between the increasing costs and the declining return earned from additional stocks. It is because striking this balance is easier to say than to do, and because it is a problem that defies solution through an intuitive understanding alone, that the new analytical concepts are necessary.

INVENTORY COSTS

This brings us face to face with the question of the costs that influence inventory policy, and the fact, noted earlier, that they are characteristically not those recorded, at least not in directly available form, in the usual industrial accounting system. Accounting costs are derived under principles developed over many years and strongly influ-

[1] See Alexander Henderson and Robert Schlaifer, "Mathematical Programing: Better Information for Better Decision-Making," HBR May–June 1954, p. 73.

enced by tradition. The specific methods and degree of skill and refinement may be better in particular companies, but in all of them the basic objective of accounting procedures is to provide a fair, consistent, and conservative valuation of assets and a picture of the flow of values in the business.

In contrast to the principles and search for consistency underlying accounting costs, the definition of costs for production and inventory control will vary from time to time—even in the same company—according to the circumstances and the length of the period being planned for. The following criteria apply:

(1) *The costs shall represent "out-of-pocket" expenditures, i.e., cash actually paid out or opportunities for profit foregone.* Overtime premium payments are out-of-pocket; depreciation on equipment on hand is not. To the extent that storage space is available and cannot be used for other productive purposes, no out-of-pocket cost of space is incurred; but to the extent that storage space is rented (out-of-pocket) or could be used for other productive purposes (foregone opportunity), a suitable charge is justified. The charge for investment is based on the out-of-pocket investment in inventories or added facilities, not on the "book" or accounting value of the investment.

The rate of interest charged on out-of-pocket investment may be based either on the rate paid banks (out-of-pocket) or on the rate of profit that might reasonably be earned by alternative uses of investment (foregone opportunity), depending on the financial policies of the business. In some cases, a bank rate may be used on short-term seasonal inventories and an internal rate for long-term, minimum requirements.

Obviously, much depends on the time scale in classifying a given item. In the short run, few costs are controllable out-of-pocket costs; in the long run, all are.

(2) *The costs shall represent only those out-of-pocket expenditures or foregone opportunities for profit whose magnitude is affected by the schedule or plan.* Many overhead costs, such as supervision costs, are out-of-pocket, but neither the timing nor the size is affected by the schedule. Normal material and direct labor costs are unaffected in total and so are not considered directly; however, these as well as some components of overhead cost do represent out-of-pocket investments, and accordingly enter the picture indirectly through any charge for capital.

DIRECT INFLUENCE

Among the costs which directly influence inventory policy are (a) costs depending on the amount ordered, (b) production costs, and (c) costs of storing and handling inventory.

Costs that depend on the amount ordered—These include, for example, quantity discounts offered by vendors; setup costs in internal manufacturing operations and clerical costs of making out a purchase order; and, when capacity is pressed, the profit on production lost during downtime for setup. Shipping costs represent another factor to the extent that they influence the quantity of raw materials purchased and resulting raw stock levels, the size of intra-plant or plant-warehouse shipments, or the size and the frequency of shipments to customers.

Production costs—Beyond setup or change-over costs, which are included in the preceding category, there are the abnormal or nonroutine costs of production whose size may be affected by the policies or control methods used. (Normal or standard raw material and direct labor costs are not significant in inventory con-

trol: these relate to the total quantity sold rather than to the amount stocked.) Overtime, shakedown, hiring, and training represent costs that have a direct bearing on inventory policy.

To illustrate, shakedown or learning costs show up wherever output during the early part of a new run is below standard in quantity or quality.[2] A cost of undercapacity operation may also be encountered—for example, where a basic labor force must be maintained regardless of volume (although sometimes this can be looked on as part of the fixed facility cost, despite the fact that it is accounted for as a directly variable labor cost).

Costs of handling and storing inventory—In this group of costs affected by control methods and inventory policies are expenses of handling products in and out of stock, storage costs such as rent and heat, insurance and taxes, obsolescence and spoilage costs, and capital costs (which will receive detailed examination in the next section).

Inventory obsolescence and spoilage costs may take several forms, including (1) outright spoilage after a more or less fixed period; (2) risks that a particular unit in stock or a particular product number will (a) become technologically unsalable, except perhaps at a discount or as spare parts, (b) go out of style, or (c) spoil.

Certain food and drug products, for example, have specified maximum shelf lives and must either be used within a fixed period of time or be dumped. Some kinds of style goods, such as many lines of toys, Christmas novelties, or women's clothes, may effectively "spoil" at the end of a season, with only reclaim or dump value. Some kinds of technical equipment undergo almost constant engineering change during their production life; thus component stocks may suddenly and unexpectedly be made obsolete.

[2] See Frank J. Andress, "The Learning Curve as a Production Tool," HBR January–February 1954, p. 87.

CAPITAL INVESTMENT

Evaluating the effect of inventory and scheduling policy upon capital investment and the worth of capital tied up in inventories is one of the most difficult problems in resolving inventory policy questions.

Think for a moment of the amount of capital invested in inventory. This is the out-of-pocket, or avoidable, cash cost for material, labor, and overhead of goods in inventory (as distinguished from the "book" or accounting value of inventory). For example, raw materials are normally purchased in accordance with production schedules; and if the production of an item can be postponed, buying and paying for raw materials can likewise be put off.

Usually, then, the raw material cost component represents a part of the out-of-pocket inventory investment in finished goods. However, if raw materials must be purchased when available (e.g., agricultural crops) regardless of the production schedule, the raw material component of finished product cost does not represent avoidable investment and therefore should be struck from the computation of inventory value for planning purposes.

As for maintenance and similar factory overhead items, they are usually paid for the year round, regardless of the timing of production scheduled; therefore these elements of burden should not be counted as part of the product investment for planning pur-

poses. (One exception: if, as sometimes happens, the maintenance costs actually vary directly with the production rate as, for example, in the case of supplies, they should of course be included.)

Again, supervision, at least general supervision, is usually a fixed monthly cost which the schedule will not influence, and hence should not be included. Depreciation is another type of burden item representing a charge for equipment and facilities already bought and paid for; the timing of the production schedule cannot influence these past investments and, while they represent a legitimate cost for accounting purposes, they should not be counted as part of the inventory investment for inventory and production planning purposes.

In sum, the rule is this: for production planning and inventory management purposes, the investment value of goods in inventory should be taken as the cash outlay made at the time of production that could have been delayed if the goods were not made then but at a later time, closer to the time of sale.

Cost of Capital Invested This item is the product of three factors: (a) the capital value of a unit of inventory, (b) the time a unit of product is in inventory, and (c) the charge or imputed interest rate placed against a dollar of invested cash. The first factor was mentioned above. As for the second, it is fixed by management's inventory policy decisions. But these decisions can be made economically only in view of the third factor. This factor depends directly on the financial policy of the business.

Sometimes businessmen make the mistake of thinking that cash tied up in inventories costs nothing, especially if the cash to finance inventory is generated internally through profits and depreciation. However, this implies that the cash in inventories would otherwise sit idle. In fact, the cash could, at least, be invested in government bonds if not in inventories. And if it were really idle, the cash very likely should be released to stockholders for profitable investment elsewhere.

Moreover, it is dangerous to assume that, as a "short-term" investment, inventory is relatively liquid and riskless. Businessmen say, "After all, we turn our inventory investment over six times a year." But, in reality, inventory investment may or may not be short-term and riskless, depending on circumstances. No broad generalization is possible, and each case must be decided on its own merits. For example:

> A great deal of inventory carried in business is as much a part of the permanent investment as the machinery and buildings. The inventory must be maintained to make operations possible as long as the business is a going concern. The cash investment released by the sale of one item from stock must be promptly reinvested in new stock, and the inventory can be liquidated only when the company is closed. How much more riskless is this than other fixed manufacturing assets?
>
> To take an extreme case, inventory in fashion lines or other types of products having high obsolescence carries a definite risk. Its value depends wholly on the company's ability to sell it. If sales are insuffi-

cient to liquidate the inventory built up, considerable losses may result.

At the other extreme, inventory in stable product lines built up to absorb short-term seasonal fluctuations might be thought of as bearing the least risk, since this type of investment is characteristically short-term. But even in these cases there can be losses. Suppose, for instance, that peak seasonal sales do not reach anticipated levels and substantially increased costs of storage and obsolescence have to be incurred before the excess inventory can be liquidated.

Finally, it might be pointed out that the cost of the dollars invested in inventory may be underestimated if bank interest rate is used as the basis, ignoring the risk-bearing or entrepreneur's compensation. How many businessmen are actually satisfied with uses of their companies' capital funds which do not earn more than a lender's rate of return? In choosing a truly appropriate rate—a matter of financial policy—the executive must answer some questions:

1. Where is the cash coming from—inside earnings or outside financing?
2. What else could we do with the funds, and what could we earn?
3. When can we get the investment back out, if ever?
4. How much risk of sales disappointment and obsolescence is really connected with this inventory?
5. How much of a return do we want, in view of what we could earn elsewhere or in view of the cost of money to us and the risk the inventory investment entails?

Investment in Facilities Valuation of investment in facilities is generally important only in long-run planning problems—as, for example, when increases in productive or warehouse capacity are being considered. (Where facilities already exist and are not usable for other purposes, and where planning or scheduling do not contemplate changing these existing facilities, investment is not affected.)

Facilities investment may also be important where productive capacity is taxed, and where the form of the plan or schedule will determine the amount of added capacity which must be installed, either to meet the plan itself or for alternative uses. In such cases, considerable care is necessary in defining the facilities investment in order to be consistent with the principles noted above: i.e., that facilities investment should represent out-of-pocket investment, or, alternatively, foregone opportunities to make out-of-pocket investment elsewhere.

CUSTOMER SERVICE

An important objective in most production planning and inventory control systems is maintenance of reasonable customer service. An evaluation of the worth of customer service, or the loss suffered through poor service, is an important part of the problem of arriving at a reasonable inventory policy. This cost is typically very difficult to arrive at, including as it does the paper work costs of rehandling back orders and, usually much more important, the effect that dissatisfaction of customers may have on future profits.

In some cases it may be possible to limit consideration to the cost of producing the needed material on overtime or of purchasing it from the outside and losing the contribution to profit which it would have made. On

the other hand, sometimes the possible loss of customers and their sales over a substantial time may outweigh the cost of direct loss in immediate business, and it may be necessary to arrive at a statement of a "reasonable" level of customer service—i.e., the degree of risk of running out of stock, or perhaps the number of times a year the management is willing to run out of an item. In other cases, it may be possible to arrive at a reasonable maximum level of sales which the company is prepared to meet with 100% reliability, being reconciled to have service suffer if sales exceed this level.

One of the uses of the analytic techniques described below and in following parts of this series is to help management arrive at a realistic view of the cost of poor service, or of the value of building high service, by laying out clearly what the cost in inventory investment and schedule changes is to achieve this degree of customer service. Sometimes when these costs are clearly brought home, even a 100% service-minded management is willing to settle for a more realistic, "excellent" service at moderate cost, instead of striving for "perfect" service entailing extreme cost.

OPTIMUM LOT SIZE

Now, with this background, let us examine in some detail one of the inventory problems which plague businessmen the most—that of the optimum size of lot to purchase or produce for stock. This happens also to be one of the oldest problems discussed in the industrial engineering texts—but this does not lessen the fact that it is one of the most profitable for a great many companies to attack today with new analytic techniques.

COMMON PRACTICES

This problem arises, as mentioned earlier, because of the need to purchase or produce in quantities greater than will be used or sold. Thus, specifically, businessmen buy raw materials in sizable quantities—carloads, or even trainloads—in order to reduce the costs connected with purchasing and control, to obtain a favorable price, and to minimize handling and transportation costs. They replenish factory in-process stocks of parts in sizable quantities to avoid, where possible, the costs of equipment setups and clerical routines. Likewise, finished stocks maintained in warehouses usually come in shipments substantially greater than the typical amount sold at once, the motive again being, in part, to avoid equipment setup and paper-work costs and, in the case of field warehouses, to minimize shipping costs.

Where the same equipment is used for a variety of items, the equipment will be devoted first to one item and then to another in sequence, with the length of the run in any individual item to be chosen, as far as is economically possible, to minimize change-over cost from one item to another and to reduce the production time lost because of clean-out requirements during change-overs. Blocked operations of this sort are seen frequently, for example, in the petroleum industry, on packaging lines, or on assembly lines where change-over from one model to another may require adjustment in feed speeds and settings and change of components.

In all these cases, the practice of

replenishing stocks in sizable quantities compared with the typical usage quantity means that inventory has to be carried; it makes it possible to spread fixed costs (e.g., setup and clerical costs) over many units and thus to reduce the unit cost. However, one can carry this principle only so far, for if the replenishment orders become too large, the resulting inventories get out of line, and the capital and handling costs of carrying these inventories more than offset the possible savings in production, transportation, and clerical costs. Here is the matter, again, of striking a balance between these conflicting considerations.

Even though formulas for selecting the optimum lot size are presented in many industrial engineering texts,[3] few companies make any attempt to arrive at an explicit quantitative balance of inventory and change-over or setup costs. Why?

For one thing, the cost elements which enter into an explicit solution frequently are very difficult to measure, or are only very hazily defined. For example, it may be possible to get a fairly accurate measure of the cost of setting up a particular machine, but it may be almost impossible to derive a precise measure of the cost of making out a new production order. Again, warehouse costs may be accumulated separately on the accounting records, but these rarely show what the cost of housing an *additional* unit of material may be. In my experience the capital cost, or imputed interest cost, connected with inventory investment never appears on the company's accounting records.

Furthermore, the inventory is traditionally valued in such a way that the true incremental investment is difficult to measure for scheduling purposes. Oftentimes companies therefore attempt to strike only a qualitative balance of these costs to arrive at something like an optimum or minimum-cost reorder quantity.

Despite the difficulty in measuring costs—and indeed because of such difficulty—it is eminently worthwhile to look at the lot size problem explicitly formulated. The value of an analytic solution does not rest solely on one's ability to plug in precise cost data to get an answer. An analytic solution often helps clarify questions of principle, even with only crude data available for use. Moreover, it appears that many companies today still have not accepted the philosophy of optimum reorder quantities from the over-all company standpoint; instead, decisions are dominated from the standpoint of some particular interest such as production or traffic and transportation. Here too the analytic solution can be of help, even when the cost data are incomplete or imperfect.

CASE EXAMPLE

To illustrate how the lot size problem can be attacked analytically—and what some of the problems and advantages of such an attack are—let us take a fictitious example. The situation is greatly oversimplified on purpose to get quickly to the heart of the analytic approach.

Elements of the Problem. Brown and Brown, Inc., an automotive parts

[3] See, for example, Raymond E. Fairfield, *Quantity and Economy in Manufacture* (New York, D. Van Nostrand Company, Inc., 1931).

supplier, produces a simple patented electric switch on long-term contracts. The covering is purchased on the outside at $0.01 each, and 1,000 are used regularly each day, 250 days per year.

The casings are made in a nearby plant, and B. and B. sends its own truck to pick them up. The cost of truck operation, maintenance, and the driver amounts to $10 per trip.

The company can send the truck once a day to bring back 1,000 casings for that day's requirements, but this makes the cost of a casing rather high. The truck can go less frequently, but this means that it has to bring back more than the company needs for its immediate day-to-day purposes.

The characteristic "saw-tooth" inventory pattern which will result is shown in Exhibit I, where 1,000 Q casings are picked up each trip (Q being whatever

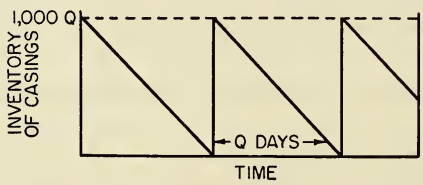


EXHIBIT I

PATTERN OF INVENTORY BALANCE

(1,000 Q casings obtained per replenishment trip; 1,000 casings used per day)

number of days' supply is obtained per replenishment trip). These are used up over a period of Q days. When the inventory is depleted again, another trip is made to pick up Q days' supply or 1,000 Q casings once more, and so on.

B. and B. estimates that the cost of storing casings under properly controlled humidity conditions is $1 per 1,000 casings per year. The company wants to obtain a 10% return on its inventory investment of $10 (1,000 times $0.01), which means that it should properly charge an additional $1 (10% of $10), making a total inventory cost of $2 per 1,000 casings per year.

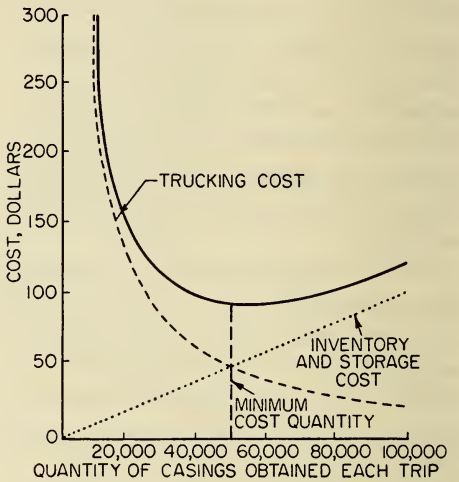


EXHIBIT II

ANNUAL COST OF BUYING, MOVING, AND STORING CASINGS COMPARED WITH REORDER QUANTITY

(Note that, in order to avoid undue complications, the inventory investment charge is made here only against the purchase price of the casings and not against the total delivery cost including transportation. Where transportation is a major component of total cost, it is of course possible and desirable to include it in the base for the inventory charge.)

Graphic Solution. Brown and Brown, Inc., can find what it should do by means of a graph (see Exhibit II)

showing the annual cost of buying, moving, and storing casings:

The broken line shows total trucking costs versus the size of the individual purchase quantity:

If 1,000 casings are purchased at a time, the total cost is \$10 times 250 trips, or \$2,500 per year.

If 10,000 casings are purchased at one time, only 25 trips need be made, for a total cost of \$250 per year.

If 100,000 casings are purchased, only 2½ trips, on the average, have to be taken each year, for a total cost of \$25.

The dotted line shows the inventory cost compared with the size of the purchased quantity:

If 10,000 casings are purchased at one time, the inventory at purchase will contain 10,000, and it will gradually be depleted until none are on hand, when a new purchase will be made. The average inventory on hand thus will be 5,000 casings. The cost per year will be \$2 times 5,000 casings, or \$10.

The solid line is the total cost, including both trucking and inventory and storage costs. The total cost is at a minimum when 50,000 casings are purchased on each trip and 5 trips are made each year, for at this point the total trucking cost and the total inventory and storage cost are equal.

The solution to B. and B.'s problem can be reached algebraically as well as graphically. Exhibit III shows how the

EXHIBIT III

EXAMPLE OF ALGEBRAIC SOLUTION OF SAME INVENTORY PROBLEM AS EXHIBIT II

The total annual cost of supplying casings is equal to the sum of the direct cost of the casings, plus the trucking cost, plus the inventory and storage cost.

Let:

T = total annual cost
b = unit purchase price, \$10 per 1,000 casings
s = annual usage, 250,000 casings
A = trucking cost, \$10 per trip
N = number of trips per year
i = cost of carrying casings in inventory at the annual rate of \$2 per 1,000 or \$0.002 per casing
x = size of an individual purchase ($x/2$ = average inventory)

Then the basic equation will be:

$$T = bs + AN + ix/2$$

The problem is to choose the minimum-cost value of x (or, if desired, N). Since x is the same as s/N, N can be expressed as s/x. Substituting s/x for N in the above equation, we get:

$$T = bs + As/x + ix/2$$

From this point on we shall use differential calculus. The derivative of total cost, T, with respect to x will be expressed as:

$$dT/dx = -As/x^2 + i/2$$

And the minimum-cost value of x is that for which the derivative of total cost with respect to x equals zero. This is true when:

$$x = \sqrt{2As/i}$$

Substituting the known values for A, s, and i:

$$x = \sqrt{2 \cdot 10 \cdot 250{,}000/.002} = 50{,}000 \text{ casings}$$

Similarly, if 100,000 casings are purchased at one time, the average inventory will be 50,000 casings, and the total inventory and storage cost will be \$100.

approach works in this very simple case.

SIMILAR CASES

The problem of Brown and Brown, Inc., though artificial, is not too far from the questions many businesses face in fixing reorder quantities.

Despite the simplifications introduced—for example, the assumption that usage is known in advance—the method of solution has been found widely useful in industries ranging from mail order merchandising (replenishing staple lines), through electrical equipment manufacturing (ordering machined parts to replenish stockrooms), to shoe manufacturing (ordering findings and other purchased supplies). In particular, the approach has been found helpful in controlling stocks made up of many low-value items used regularly in large quantities.

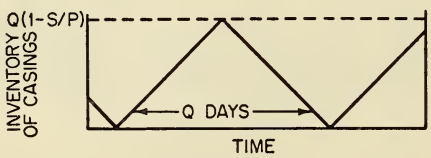


EXHIBIT IV

INFLUENCE OF PRODUCTION AND SALES RATE ON PRODUCTION CYCLE INVENTORY

A number of realistic complications might have been introduced into the Brown and Brown, Inc., problem. For example:

In determining the size of a manufacturing run, it sometimes is important to account explicitly for the production and sales rate. In this case, the inventory balance pattern looks like Exhibit IV instead of the saw-tooth design in Exhibit I. The maximum inventory point is not equal to the amount produced in an individual run, but to that quantity less the amount sold during the course of the run. The maximum inventory equals $Q\ (1 - S/P)$, where Q is the amount produced in a single run, and S and P are the daily sales and production rates respectively.

This refinement can be important, particularly if the sales rate is fairly large compared with the production rate. Thus, if the sales rate is half the production rate, then the maximum inventory is only half the quantity made in one run, and the average inventory equals only one-fourth the individual run quantity. This means that substantially more inventory can be carried—in fact, about 40% more.

When a number of products are made on a regular cycle, one after another, with the sequence in the cycle established by economy in change-over cost, the total cycle length can be obtained in the same way as described above. Of course, it sometimes happens that there is a periodic breach in the cycle, either to make an occasional run of a product with very low sales or to allow for planned maintenance of equipment; the very simple run-length formulas can be adjusted to allow for this.

Other kinds of costs can also be included, such as different sorts of handling costs. Or the inventory cost can be defined in such a way as to include transportation, obsolescence, or even capital and storage cost as part of the unit value of the product against which a charge for capital is made. When a charge for capital is included as part of the base value in computing the cost of capital, this is equivalent to requiring that capital earnings be compounded; this can have an important bearing on decisions connected with very low volume

items which might be purchased in relatively large, long-lasting quantities.

Complications such as the foregoing, while important in practice, represent changes in arithmetic rather than in basic concept.

SIGNIFICANT CONCLUSIONS

When the analytic approach is applied to Brown and Brown's problem and similar cases, it reveals certain relationships which are significant and useful to executives concerned with inventory management:

(1) *The appropriate order quantity and the average inventory maintained do not vary directly with sales.* In fact, both of these quantities vary with the square root of sales. This means that with the same ordering and setup cost characteristics, the larger the volume of sales of an item, the less inventory per unit of sales is required. One of the sources of inefficiency in many inventory control systems is the rigid adoption of a rule for ordering or carrying inventory equivalent to, say, one month's sales.

(2) *The total cost in the neighborhood of the optimum order quantity is relatively insensitive to moderately small changes in the amount ordered.* Exhibit II illustrates this proposition. Thus, all that is needed is just to get in the "right ball park," and a good answer can be obtained even with fairly crude cost data. For example, suppose the company had estimated that its total cost of holding 1,000 casings in inventory for a year was $1 when it actually was $2 (as in our illustration). Working through the same arithmetic, the company would have arrived at an optimum order quantity of 70,000 casings instead of 50,000. Even so, the total cost would have been (using the correct $2 annual carrying cost):

3.6 trips per year @ $10	= $36
35,000 casings average inventory @ $0.002	= 70
Total annual cost	= $106

Thus, an error of a factor of 2 in one cost results in only a 6% difference in total cost.

In summary, Brown and Brown's problem, despite its oversimplification, provides an introduction to the analytic approach to inventory problems.

In particular, it illustrates the first essential in such an approach—i.e., defining an inventory function. In this case the function is to permit purchase or manufacture in economical order quantities or run lengths; in other cases it may be different. The important point is that this basic function can be identified wherever it may be found—in manufacturing, purchasing, or warehouse operation.

The only way to cut inventories is to organize operations so that they are tied more closely together. For example, a company can cut its raw materials inventory by buying in smaller quantities closer to needs, but it does so at a cost; this cost results from the increased clerical operations needed to tie the purchasing function more closely to manufacturing and to keep it more fully informed of manufacturing's plans and operations. The right inventory level is reached when the cost of maintaining any additional inventory cushion offsets the saving that the additional inventory earns by permitting the plant to operate in a somewhat less fully organized fashion.

B. and B.'s problem also illustrates problems and questions connected with

defining and making costs explicit. The inventory capital cost is usually not found on a company's books, but it is implied in some of the disagreements over inventory policy. Here, again, bringing the matter into the open may help each side in a discussion to recognize its own and the others' hidden assumptions, and thus more quickly to reach a common agreement.

✦✦✦✦✦✦✦✦✦✦✦✦✦ GUIDES TO INVENTORY POLICY: *problems of uncertainty*

JOHN F. MAGEE

Marketing and production executives alike have an immediate, vital interest in safety stocks. In these days of strong but often unpredictable sales, safety stocks afford, for the factory as well as for the sales office, a method of buying short-term protection against the uncertainties of customer demand. They are the additional inventory on hand which can be drawn upon in case of emergency during the period between placement of an order by the customer and receipt of the material to fill the order. However, in practice their potentials are often needlessly lost.

One reason for the failure is a very practical one. Because safety stocks are designed to cope with the uncertainties of sales, they must be controlled by flexible rules so that conditions can be met as they develop. But sometimes the need for flexibility is used as an excuse for indefiniteness: "We can't count on a thing; we have to play the situation by ear." And, in any sizable organization, when people at the factory level start "playing it by ear," one can be almost sure that management policy will not be regularly translated into practice.

Our studies have shown that the methods used by existing systems in industry often violate sound control concepts. The economy of the company is maintained, in the face of instability and inefficiency in the inventory control system, only because of constant attention, exercise of overriding common sense, and use of expediting and other emergency measures outside the routine of the system.

Actually, it is possible to have inventory controls which are not only flexible but also carefully designed and explicit. But the task needs special analytical tools; in a complicated business it defies common-sense judgment and simple arithmetic. Methods must be employed to take direct account of uncertainty and to measure the response characteristics of the system and relate them to costs. Such methods are the distinctive mark of a really modern, progressive inventory control system.

Harvard Business Review, *March–April 1956, 34:2, 103–116.*

Here are some of the points which I shall discuss in this article:

Basically, there are two different types of inventory replenishment systems designed to handle uncertainty about sales—*fixed order*, commonly used in stockrooms and factories, as in bins of parts or other materials; and *periodic reordering*, frequently used in warehouses for inventories involving a large number of items under clerical control. While the two are basically similar in concept, they have somewhat different effects on safety stocks, and choice of one or the other, or some related variety, requires careful consideration. Certain factors which should be taken into account in the choice between them will be outlined.

The fundamental problem of setting safety stocks under either system is balancing a series of types of costs which are not found in the ordinary accounting records of the company—costs of customer service failure, of varying production rates (including hiring and training expenses), of spare capacity, and others. Often specialists can find the optimum balance with relatively simple techniques once the cost data are made explicit. However, part of the needed data can come *only from top management*. For example, the tolerable risk of service failure is generally a policy decision.

The specific problem of inventory control, including production scheduling, varies widely from company to company. Where finished items can be stocked, the important cost factors to weigh may be storage, clerical procedures, setup, supervision, etc. But where finished items cannot be stocked, the problem is one of setting capacity levels large enough to handle fluctuating loads without undue delay, which involves the cost of unused labor and machines. Despite the great variety of situations that are possible, specific mathematical approaches and theories are available for use in solving almost any type of company problem.

Both to illustrate the various techniques and by way of summary, a hypothetical case will be set forth where a company moved through a series of stages of inventory control. Significantly, the final step brought a large reduction in stocks needed for efficient service and also a great reduction in production fluctuations. Out of the range of this company's experience, other managements should be able to get some guidance as to what is appropriate for their own situations.

BASIC SYSTEMS

Like transit stocks and lot-size stocks (discussed specifically in the previous article in this series [and book] [1]) and also anticipation stocks (to be taken up in a subsequent article), safety stocks "decouple" one stage in production and distribution from the next, reducing the amount of over-all organization and control needed.

But the economies of safety inventories are not fairly certain and immediate. The objective is to arrive at a reasonable balance between the costs of the stock and the protection obtained against inventory exhaustion. Since exhaustion becomes less likely as the safety inventory increases, each additional amount of safety inventory characteristically buys relatively less protection. The return from increasing inventory balances therefore diminishes

[1] John F. Magee, "Guides to Inventory Policy: I. Functions and Lot Sizes," HBR January–February 1956, p. 49. Reprinted in this book.

rapidly. So the question is: How much additional inventory as safety stock can be economically justified?

To answer this question we need to look at the two basic systems of inventory replenishment to handle uncertainty about sales and see how they produce different results.

FIXED ORDER

Under any fixed order system—the old-fashioned "two-bin" system or one of its modern varieties—the same *quantity* of material is always ordered (a binful in the primitive system), but the *time* an order is placed is allowed to vary with fluctuations in usage (when the bottom of one bin is reached). The objective is to place an order whenever the amount on hand is just sufficient to meet a "reasonable" maximum demand over the course of the lead time which must be allowed between placement of the replenishment order and receipt of the material.

Where the replenishment lead time is long (e.g., three months) compared with the amount purchased at each order (e.g., a one-month supply), there are presumably some purchase orders outstanding all the time which, on being filled, will help replenish the existing inventory on hand. In such cases, of course, the safety stocks and reorder points should be based upon both amount on hand and on order. Where, on the other hand, the lead time is short compared with the quantity ordered, as in most factory two-bin systems, the amount on hand and the total on hand and on order are in fact equivalent at the time of reordering.

The key to setting the safety stock is the "reasonable" maximum usage during the lead time. What is "reasonable" depends partly, of course, on the nature of short-term fluctuations in the rate of sale. It also depends—and here is where the top executive comes foremost into the picture—on the risk that management is prepared to face in running out of stock. What is the level of sales or usage beyond which management is prepared to face the shortages? For example:

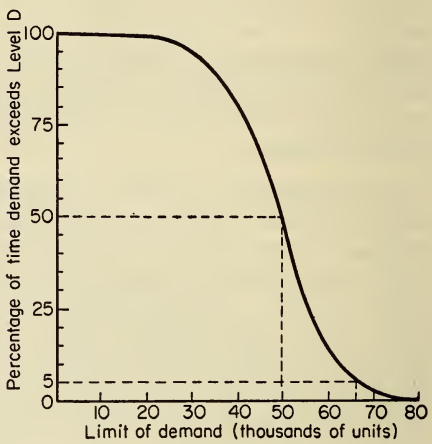


EXHIBIT I

BROWN AND BROWN'S SAFETY STOCK

In Exhibit I, continuing the hypothetical case of Brown and Brown, Inc., discussed in the first article in this series,[2] the curve shows the number of weeks in which the demand for casings may be expected to equal or exceed any specified level. (Such a curve could be roughly plotted according to actual experience modified by such expectations or projections as seem warranted; refinement can be added by the use of mathematical analysis when such precision seems desirable.)

Now, if it takes B. and B. a week to re-

[2] Ibid., p. 57.

plenish its stocks and the management wishes to keep the risk of running out of stock at a point where it will be out of stock only once every 20 weeks, or 5% of the time, then it will have to schedule the stock replenishment when the inventory of casings on hand drops to 66,000 units. Since the expected or average weekly usage is 50,000 units, the safety stock to be maintained is 16,000 (making a total stock of 66,000).

This example, of course, assumes a single, rather arbitrary definition of what is meant by risk or minimum acceptable level of customer service. There are a number of ways of defining the level of service, each appropriate to particular circumstances. One might be the total volume of material or orders delayed; another, the number of customers delayed (perhaps only in the case of customers with orders exceeding a certain size level), still another the length of the delays. All of these definitions are closely related to the "probability distribution" of sales —i.e., to the expected pattern of sales in relation to the average.

Cost of Service Failure. It is easy enough to understand the principle that setting a safety stock implies some kind of a management decision or judgment with respect to the maximum sales level to be allowed for, or the cost of service failure. But here is the rub: service failure cost, though real, is far from explicit. It rarely, if ever, appears on the accounting records of the company except as it is hidden in extra sales or manufacturing costs, and it is characteristically very hard to define. What is new in inventory control is not an accounting technique for measuring service cost but a method of self-examination by management of the intuitive assumptions it is making. The progressive company looks at what it is in fact assuming as a service-failure cost in order to determine whether the assumed figure is anywhere near realistic.

For example, characteristically one hears the policy flatly stated: "Back orders are intolerable." What needs to be done is to convert this absolute, qualitative statement into a quantitative one of the type shown in Exhibit II. Here we see the facts which might be displayed for the management of a hypothetical company to help it decide on a customer service policy:

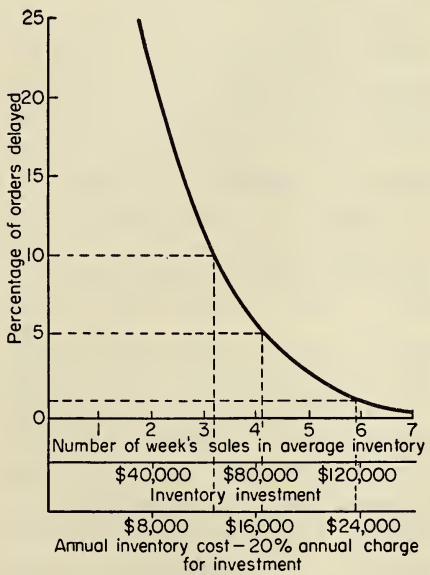


EXHIBIT II

RELATION BETWEEN SAFETY STOCKS AND ORDER DELAY

To get a 90% level of customer service (i.e., to fill 90% of the orders immedi-

ately), a little over three weeks' stock must be carried—an investment of $64,000 with an annual carrying cost of $12,800.

Filling another 5% of orders immediately, thereby increasing the service level to 95%, would mean about one week's more stock, with an extra annual cost of $3,800.

Filling another 4% immediately (a 99% service level) would cost an extra $7,400 per year.

At each point the management can decide whether the extra cost is justified by the improved service. Thus, the chart becomes a device for comparing policies on service and inventories for consistency and rationality.

PERIODIC REORDERING

The periodic reordering system of inventory replenishment—the other basic approach to handling uncertainty about demand—is very popular, particularly where some type of book inventory control is employed and where it is convenient to examine inventory stocks on a definite schedule. The idea underlying all varieties of this system is to look at stocks at fixed *time* intervals, and to vary the order *amount* according to the usage since the last review.

The problem is that many seemingly similar ways of handling a cyclical ordering system may have hidden traps. A typical difficulty is instability in reordering habits and inventory levels caused by "overcompensation"; that is, by attempting to outguess the market and assuming that high or low sales at one point, actually due to random causes, indicate an established trend which must be anticipated. For example:

An industrial abrasives manufacturer found himself in a characteristic state of either being out of stock or having too much stock, even though his inventory control procedures were, at least judging by appearances, logically conceived. The procedures worked as follows: Each week the production scheduling clerk examined the ledger card on each item, and each month he placed a replenishment order on the factory based on (a) the existing finished stock on hand in the warehouse, (b) a replenishment lead time of six weeks, and (c) a projection for the coming two-month period of the rate of sales during the past two-month period.

The manufacturer blamed the instability of his market and the perversity of his customers for the difficulties he faced in controlling inventory, when in fact the seemingly logical reorder rule he had developed made his business behave in the same erratic fashion as a highly excitable and nervous driver in busy downtown traffic. The effects of sales fluctuations tended to be multiplied and passed on to the factory. *No use was made of inventories—especially safety stock—to absorb sales fluctuations.*

The most efficient and stable reorder scheme or rule has a very simple form:

A forecast or estimate of the amount to be used in the future is made for a period equal to the delivery lead time plus one reorder cycle. Then an order is placed to bring the total inventory on hand and on order up to the total of the amount forecast for the delivery lead and cycle times, plus a standard allowance for safety stock. Under such a scheme, the average inventory expected to be on hand will be the safety balance plus one-half the expected usage during a reorder cycle.

Note the contrast between this scheme and that used by the abrasives

manufacturer. Here inventories are used to "decouple" production and sales. An upward fluctuation in sales is "absorbed" at the warehouse; it is not passed on to the plant until later (if at all). Many companies subscribe to this plan wholeheartedly in principle but only halfheartedly in practice. A common tendency, for instance, is to make the forecast but then, if sales increase, to revise it upward and transmit the increase back to the plant. The whole value of a safety stock based on a balancing of the costs of running out and the costs of rush orders to production is thus lost.

Readers may recognize the application here of servo theory, the body of concepts (including feedback, lags or reaction times, type of control, and the notion of stability) developed originally by electrical engineers in designing automatic or remotely controlled systems.[3] An inventory system, though not a mechanical device, is a control system and as a consequence is subject to the same kinds of effects as mechanical control systems and can be analyzed using the same basic concepts.

CHOICE OF SYSTEM

Each system of reordering inventories has its own advantages. Here are the conditions under which the fixed order system is advantageous:

Where some type of continuous monitoring of the inventory is possible, either because the physical stock is seen and readily checked when an item is used or because a perpetual inventory record of some type is maintained.

Where the inventory consists of items of low unit value purchased infrequently in large quantities compared with usage rates; or where otherwise there is less need for tight control.

Where the stock is purchased from an outside supplier and represents a minor part of the supplier's total output, or is otherwise obtained from a source whose schedule is not tightly linked to the particular item or inventory in question; and where irregular orders for the item from the supplier will not cause production difficulties.

For example, the fixed order system is suitable for floor stocks at the factory, where a large supply of inexpensive parts (e.g., nuts and bolts) can be put out for production workers to draw on without requisitions, and where a replenishment is purchased whenever the floor indicates the supply on hand has hit the reorder point.

By contrast, the periodic reordering system is useful under these conditions:

Where tighter and more frequent control is needed because of the value of the items.

Where a large number of items are to be ordered jointly, as in the case of a warehouse ordering many items from one factory. (Individual items may be shipped in smaller lots, but the freight advantages on large total shipments can still be obtained.)

Where items representing an important portion of the supplying plant's output are regularly reordered.

In general, since safety stocks needed vary directly with the length of the period between orders, the periodic system is less well suited where the cost of ordering and the low unit value of

[3] See H. J. Vassian, "Application of Discrete Variable Servo Theory to Inventory Control," *Journal of the Operations Research Society of America*, August 1955, p. 272.

the item mean infrequent large orders.

It should be noted that modifications of the simplest fixed order system or intermediates between the fixed order system and the periodic reordering system are also possible and very often useful; they can combine the better control and cost features of each of the "pure" schemes. For example:

One type of scheme often useful—the "base stock" system—is to review inventory stocks on a periodic basis but to replenish these stocks only when stocks on hand and on order have fallen to or below some specified level. When this happens, an order is placed to bring the amount on hand and on order up to a specified maximum level.

The choice of frequency of review and the minimum and maximum inventory points can be determined by analysis similar to that used for the other systems, but precautions must be taken—such as that stocks on order must always be counted when reorder quantities are figured—in order to avoid problems of instability and oscillation which can easily creep into rules that are apparently sound and sensible.

Interaction among Factors. As mathematical analysis will indicate, the safety stock, reorder quantity, and reorder level are not entirely independent under either the fixed order or the periodic reordering system (or any combination thereof):

Where the order amount is fixed, the safety stock is protection against uncertainty over the replenishment time (measured by the reorder level). But it is the size of the order amount that determines the frequency of exposure to risk. With a given safety level, the bigger the order placed, the less frequently will the inventory be exposed to the possibility of run-out and the higher will be the level of service.

Where inventories are reordered on a periodic time cycle, the uncertainty against which safety stocks protect extends over the *total* of the reorder period and replenishment time. But here it is the length of the reordering cycle that determines the risk. The shorter the period and the closer together the reorders, the less will be the chance of large inventory fluctuations and, as a consequence, the less will be the size of safety stock required in order to maintain a given level of service.

The interaction among the frequency of reorder, the size of reorder, and safety stocks is often ignored as being unimportant, even in setting up fairly sophisticated inventory control schemes (although the same companies readily consider the *lot-size* problem in relation to the other factors). In many cases this may be justifiable for the purpose of simplifying inventory control, particularly methods for adjusting reorder quantities and safety stocks to changing costs and sales. On the other hand, cases do arise from time to time where explicit account must be taken of such interactions so that an efficient system may be developed.

Note, too, that the factors governing the choice of any reorder scheme are always changing. Therefore, management should provide for routine review of the costs of the system being used, once a year or oftener, so that trends can be quickly identified. Also, control chart procedures, like simple quality control methods, should be used to spot "significant" shifts in usage rates and in the characteristics of customer demand (fluctuations, order size, fre-

quency of order, etc.). Schemes for checking such matters each time a reorder point is crossed are easily incorporated in the programs of automatic data-handling systems used for inventory control; they can also be applied to manual systems, but less easily and hence with some temptation to oversimplify them dangerously.

PRODUCTION SCHEDULING

Now let us turn to the important relationships between safety stocks and production. The safety stock affects, and is affected by, production run cycles, production "reaction times," and manufacturing capacity levels.

SETTING CYCLE LENGTHS

In production cycling problems, as in periodic reordering, the longer the run on each product, the longer one must wait for a rerun of that product; therefore, a larger safety stock must be maintained as protection. Shorter, more frequent runs give greater flexibility and shorter waiting periods between runs, and thus lower safety inventory requirements. Also, again the interaction between factors must be taken into account. For example:

A chemical company arrived at production run cycles for a set of five products going through the same equipment on the basis of only setup costs and cycle inventories (e.g., lot-size inventories), ignoring the interaction between cycle length and safety stocks. It found that on this basis an over-all product cycle of approximately 20 days, or one production month, appeared optimum, allowing 4 days per product on the average. However, when the problem was later re-examined, it was discovered that the uncertainty introduced by long lead times was so great that the over-all product cycle could in fact be economically cut back to less than 10 days. Doubling setup costs would be more than offset by savings in inventory and storage costs resulting from a reduction in the needed safety stocks.

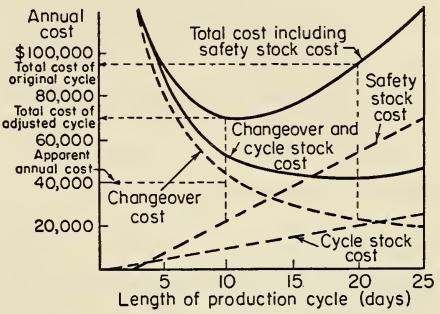


EXHIBIT III

INFLUENCE OF SAFETY STOCKS ON CHOICE OF AN OPTIMUM PRODUCTION CYCLE

Exhibit III illustrates the cost characteristics found to exist. The three *dashed* lines show separately the annual costs of changeovers, carrying cycle inventories, and carrying safety stocks, compared with the length of the individual production cycle. Adding together only the first two costs leads to the lower of the *solid* lines. This is at a minimum when the production cycle is 20 days long, indicating an apparent annual cost of $40,000. However, if *all* costs are included (the *solid* line at the top), the total annual cost on a 20-day cycle is $95,000. On this basis total costs are at a minimum when the cycle is 10 days long—only $70,000. This means a saving of $25,000 annually on the products in question.

SETTING PRODUCTION LEVELS

Safety stocks give only short-term protection against sales uncertainty. If stocks are being replenished from production, the effectiveness of over-all control depends also on the ability to restore them in case of depletion.

If total demand varies, the ability to restore stocks depends, in turn, on the ability of the production facilities to react to chance fluctuations. In order to get low inventories, the process must have fast reactions properly controlled or (equivalently) in some cases large "capacity." If reactions are slow or limited, inventories must be large, and the inventory in effect serves another type of protective function, namely, protection of production rate or capacity from the stresses of demand fluctuation. To illustrate the kind of situation where this may be true:

Changes in the throughput rate of chemical processing equipment may be slow and difficult or expensive.

The output level of an assembly line operation may depend on the number of stations that are manned, or the number of shifts working. Some time may be required to change the production rate by changing the number of stations manned at each point along the line.

The production output of a job-shop operation may be influenced by the rate at which new workers can be hired and trained, or the cost of making changes in the manning level by bringing in new untrained workers or laying off people.

How fast should production operations respond to sales fluctuations, and to what extent should these fluctuations be absorbed by means of inventory? The costs of warehousing and cash investment in inventory need to be balanced against the costs of changing production rates or building excess capacity into the production system.

The actual cost of making out schedules, which depends on the frequency with which they are made and the degree of precision required, also should be considered, as well as the speed of reaction of production which is physically possible (e.g., the employee training time). When these costs are made explicit, management may find itself having to balance conflicting objectives. To illustrate:

A metal fabricator making a wide line of products to order attempted to provide immediate service to customers. He found that on the average his departments needed a substantial excess of labor over the normal requirements of the jobs flowing through, and this excess was essentially idle time. On the other hand, when he attempted to cut the excess too thin, backlogs began to build up. He had to weigh his desire to get the lead time down against the costs of excess unused labor.

Ordinarily we want to avoid passing back the full period-to-period sales fluctuation by making corresponding changes in the size of orders placed on production because it is uneconomical. What we can do instead is to:

1. Set the production level in each period equal to anticipated needs over the lead time plus the scheduling period not already scheduled, plus or minus *some fraction* of the difference between desired and actual inventory on hand.

2. Alternatively, change the existing production level or rate by *some fraction* of the difference between the existing rate and the rate suggested by the simple re-

order rule (i.e., that an order be placed in each period equal to the anticipated requirements over the lead time plus the scheduling period, plus or minus the difference between desired and actual inventory on hand and on order).

Each of these alternatives is useful in certain types of plants, depending on whether the cost of production fluctuations comes primarily from, say, overtime and undertime (work guarantee) costs or from hiring, training, and layoff costs. Each in appropriate circumstances will lead to smoother production, at the expense of extra inventory to maintain the desired level of service.

When the different costs involved are identified and measured, mathematical techniques can be used to show the effect that varying the numbers in the rule (in particular, the size of the *fraction* used) has on inventory and production expense and to arrive at an economical balance between the needs of marketing and manufacturing. These two rules are expressions of servo theory, like that referred to earlier in connection with inventory. Here it may be worthwhile to see in some working detail how the theory can be applied mathematically:

The first rule can be stated as follows:

$$P_t = \sum_{k=0}^{T} F_{t+k} - \sum_{k=1}^{T} P_{t-k} + k(I_n - I_t);\ k \leq 1$$

P_t is the amount scheduled for production in period i, F_t is the forecast requirements for period i, I_o is the desired inventory, I_t is the actual opening inventory on hand in period i, and k is the response number which indicates what fraction of the inventory error or production rate departure is to be accounted for each period.

The fluctuations in inventory resulting from a choice of k in the first rule can be expressed as a function of the fluctuations in sales about the forecast, as follows (if fluctuations from month to month are not correlated):

$$\sigma_I = \sqrt{\frac{T(2k - k^2) + 1}{2k - k^2}}\ \sigma_F$$

where σ_I is the standard deviation of inventory levels, and σ_F is the standard deviation of actual sales about forecast sales each period. Similarly, the production rate variations resulting from any choice of k can be expressed as:

$$\sigma_P = \sqrt{\frac{k}{2 - k}}\ \sigma_F$$

The influence of the choice of a response number, k, on the standard deviation of inventories and on the standard deviation of production rates under the first type of rule is shown in Exhibit IV. Frequently the costs of production fluctuations are more or less directly proportional to the standard deviation of fluctuations in the production rate, a measure of the amount of change in production level which can be expected to occur. On the other hand, the normal inventory level, the average level expected, must be set large enough so that even with expected inventory fluctuations, service failures will not occur excessively. This means that the larger the standard deviation in inventory levels, the larger must be the normal level, generally in proportion. Therefore, one can "buy" production flexibility with larger inventories, and vice versa, with the particular costs in the process concerned determining the economical balance.[4]

[4] See H. J. Vassian, op. cit. See also Charles C. Holt, Franco Modigliani, and Herbert A. Simon, "A Linear Decision Rule for Produc-

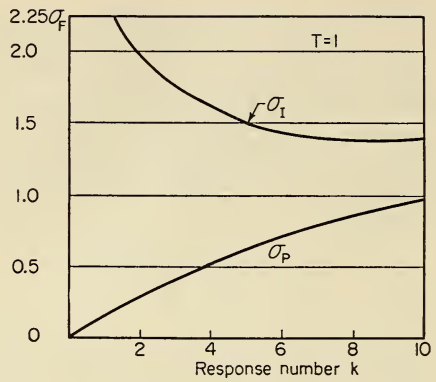


EXHIBIT IV

EFFECT OF RESPONSE NUMBER k ON VARIATIONS IN INVENTORY AND PRODUCTION RATE

The second rule can be worked through similarly. Here P^* is the changed amount scheduled for production, and the rule can be stated as follows:

$$P_t^* = P_{t^*-1} + k(P_t - P_{t^*-1}); \; k \leq 1 = (1-k)P_{t^*-1} + kP_t$$

where

$$P_t = \sum_{k=0}^{T} F_{t+k} - \sum_{k=1}^{T} P_{t^*-k} + (I_o - I_t)$$

SETTING CAPACITY LEVELS

In some cases—particularly where output cannot be stocked easily—the problem of controlling the production level is not so much one of adjusting the level to respond to fluctuations in demand, as of setting the capacity of the plant or operation at a high enough level to permit demand fluctuations to be absorbed without excessive delay.

tion and Employment Scheduling," *Management Science*, October 1955, for another approach to this problem under different cost conditions.

If the capacity is set equal only to the desired average rate, fluctuations in demand about this desired rate must either be absorbed by inventories or by orders piling up in a backlog. To illustrate:

The telephone companies have recognized for many years that telephone exchanges must be built with greater capacity than is required to handle the average load, in order to keep lines of waiting subscribers within reasonable levels.

Pile-ups often occur around the checkout booths of cafeterias or the ticket windows in railroad stations. Customers are eventually taken care of, but capacity is so close to average requirements, in some cases, that long waiting lines can be built up as a result of customers arriving at random in small bunches.

The problem of specifying the number of workmen to tend semiautomatic machinery or the capacity of docks to service freighters is complicated by the fact that the units require service more or less at random, so that again there can easily develop an accumulation of units awaiting service if personnel are not immediately available.

A theory of such processes is growing; it is known as waiting-line theory. This is really a branch of probability theory, and is itself a whole body of mathematical techniques and explicit concepts providing a mathematical framework within which waiting-line and similar problems can be studied.[5]

Some examples of applications in

[5] A technical discussion of waiting-line theory and related applications can be found in W. Feller, *An Introduction to Probability Theory and Its Applications* (New York, John Wiley & Sons, Inc., 1950), Chapter 17.

production scheduling are: flow of orders through departments in a job shop; flow of items through the stages in an assembly line; clerical processing of orders for manufacture or shipping; filling orders in a warehouse or stockroom; and setting up shipping or berth facilities to handle trucks or other transport units. In each case, fairly well-fixed crews or facilities have to be set up for handling fluctuating orders or items quickly, avoiding delays in service. A balance between the cost of extra personnel or facilities and delays in taking care of demand is needed.

In applying waiting-line theory to such problems, the flow of orders or demand for goods can be considered as a demand for service, analogous to subscriber cost in a telephone exchange. Orders are handled by one or more processing stations, analogous to telephone trunk lines. When the order or unit is produced, the processing station is free to take on the next order in line, as when a call is completed through the exchange. For example:

A wholesale merchandise house planned its order-handling and order-filling activities in advance of peak sales. The company, selling consumer merchandise to a large group of retail dealers, had grown rapidly and in mid-summer had looked forward to serious congestion, delayed orders, and lost customers when the Christmas peak hit. An analysis based on waiting-line theory outlined staff and space requirements to meet the forecast load, showed what jobs were the worst potential bottle necks, and revealed, incidentally, how the normally inefficient practice of assigning two persons to "pick" one order could in this case help avoid tie ups and save space during the critical sales peak.

STAGES OF CONTROL

The choice and use of appropriate techniques for inventory control is not a simple matter. It takes a good deal of research into sales and product characteristics, plus skill in sensing which of many possible approaches are likely to be fruitful.

To describe these techniques, I shall take a case illustration. This case is drawn from a great deal of business experience, but in order to keep the detail and arithmetic within manageable proportions without distorting the essential points, I have simplified and combined everything into one fictional situation.

Any of the stages of the company's progress toward more efficient inventory management—from the original to the final—might be found to exist in the inventory control practices of a number of sizable companies with reputations for progressive and efficient management. These stages of advancement in the refinement of inventory control should not be used to compare the inventory system of one company or division with that of another, for the reasons just mentioned; but they may prove helpful to management in answering the questions, "Where are we now?" and "What could we do better?"

Briefly, the case situation is as follows:

One division of the Hibernian Bay Company makes and sells a small machine part. Sales run slightly over 5,000 units annually, and the price is $100 apiece. Customers are supplied from four branch stock

points scattered about the country, which in turn are supplied by the factory warehouse. The machining and assembly operations are conducted in a small plant, employing largely semiskilled female help. The level of production can be changed fairly rapidly but at the cost of training or retraining workers, personnel office expenses, and increased inspection and quality problems. The division management has almost complete autonomy over its operations, although its profit records are closely scrutinized at headquarters in Chicago.

Originally the factory and branch warehouse stocking practices were haphazard and unsatisfactory. In total, nearly four months' stock was carried in branches, in the factory warehouse, or in uncompleted production orders. A stock clerk in each branch who watched inventories and placed reorders on the factory warehouse was under pressure to be sure that stocks were adequate to fill customer orders. The factory warehouse reorder clerk in turn watched factory stocks and placed production orders. Production runs or batches were each put through the plant as a unit. Fluctuations in production, even with apparently sizable stocks on hand, caused the management deep concern.

SERVICE IMPROVED

The management decided to try to improve inventory practices and appointed a research team to study the problem. The team suggested using "economical order quantities" for branch orders on the factory warehouse and warehouse orders on production, as a basis for better control. The steps followed were:

The research team suggested that the formula for determining the economical order quantity was $x = \sqrt{2As/i}$, where A = fixed cost connected with an order (setup of machines, writing order, checking receipts, etc.), i = annual cost of carrying a unit in inventory, s = annual movement, and x = "economical order quantity."

The team found that each branch sold an average of 25 units a week, or 1,300 per year; that the cost of a branch's placing and receiving an order was \$19 (\$6 in clerical costs at the branch and factory, \$13 in costs of packing and shipping goods, receiving, and stocking): that annual inventory carrying costs in the branches were \$5 per unit, based on a desired 10% return on incremental inventory investment. The reorder quantity for each branch was computed as $\sqrt{2 \cdot \$19 \cdot 1{,}300/\$5} =$ 100 unit reorder quantity.

A system was set up where each branch ordered in quantities of 100, on the average, every four weeks. On this basis, without further action, each branch would have had an average inventory of one-half a reorder quantity, or 50 units. (The books would show 75 units, since stock in transit from factory warehouse to branch was also charged to the branch, and with average transit time of one week this would average 25 units.)

The next step was to provide for enough to be on hand when a reorder was placed to last until the order was received. While the average transit time was one week, experience showed that delays at the factory might mean an order would not be received at the branch for two weeks. So sales for two weeks had to be covered.

Statistical analysis showed that sales in any one branch over two weeks could easily fluctuate from 38 units to 62 units and could conceivably go as high as 65–70. The management decided that a 1% chance of a branch running out of stock before getting an order would be adequate.

Calculations then indicated that the

maximum reasonable two-week demand to provide for would be 67. (The statistical basis was that sales fluctuate about the average at random; that fluctuations in the various branches are independent of one another; and that the standard deviation is $\sqrt{st}$ where s = sales rate, and t = length of individual time period.)

The branches therefore were instructed to order 100 units whenever the stock on hand and on order was 67 or less. This gave an inventory in each branch made up on the average as follows:

Safety stock	42	(order point, 67, less normal week's usage, 25)
Order cycle stock	50	(one half 100-unit order)
In transit	25	(one week's sales)
Total	117	or 4.7 weeks' sales

The resulting behavior of the reorder system is shown in Exhibit v—both as it would be presumed in theory and as it actually turned out. Although the actual performance was much less regular than presumed, the two compare fairly well—testimony to the soundness of the procedure.

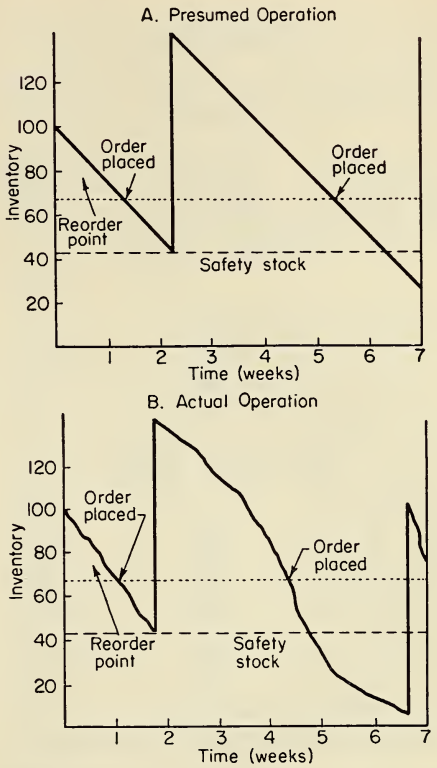


EXHIBIT V

ECONOMICAL REORDER SYSTEM OF A BRANCH WAREHOUSE

APPLICATION AT THE FACTORY

At the factory warehouse end, the "economical order quantity" scheme worked as follows:

The cost of holding a unit in inventory was \$3.50 per year (at 10% return on investment); the cost of placing an order and setting up equipment for each order was \$13.50; and, of course, a total of 5,200 units was made each year. These indicated that each production order should be for $\sqrt{2 \cdot \$13.50 \cdot 5{,}200/\$3.50} = 200$ units.

Factory processing time was two weeks; it would take two weeks for each order to reach the warehouse. The warehouse would need to place its replenishment order on the factory when it had enough on hand or on order to fill maximum reasonable demand during the next two weeks.

On the average, the factory warehouse would receive one order a week from the branches (one every four weeks from each of four branches) under the new branch reorder system. In fact, because of the fluctuations in branch sales described before, it was found that orders on the factory warehouse fluctuated substantially in any two-week period (see Exhibit vi).

EXHIBIT VI

FLUCTUATIONS OF ORDERS ON FACTORY WAREHOUSE

Number of branch orders	*Number of items ordered*	*Percentage of weeks*
A. Weekly Periods		
0	0	37%
1	100	37
2	200	18
3	300	6
4+	400+	2
B. Biweekly Periods		
0	0	13%
1	100	27
2	200	27
3	300	18
4	400	9
5	500	4
6	600	1
7+	700+	1

It was agreed that to give branches service adequate to maintain their own service, stocks at the factory warehouses would have to be high enough to fill demand 99% of the time, i.e., a replenishment order would have to be placed when 600 units were on hand. This meant a safety stock of 600 units minus 200 (normal usage), or 400 units. Cycle stock averaged half a run, or 100 units, and stock in process an additional half run, or 100 units. Total factory stock, then, was:

Cycle stock	100 units
Stock in process	100
Safety stock	400
Total	600 units

EXHIBIT VII gives a picture of the apparent costs of the "economical order" system. The stock of 1,068 units equaled less than 11 weeks' sales, a fairly substantial reduction, and the management felt that it had a better control, since clerical procedures were set up to adapt readily to any changes in inventory charges (currently 10% per year) or service level requirements the management might choose to make.

PRODUCTION STABILIZED

But the factory still had problems. On the average, the warehouse would place one production order every two

EXHIBIT VII

COSTS OF REORDER SYSTEM

	Number	*Cost each*	*Annual cost*
Inventory			
Factory	600 units	$3.50/year	$2,100
4 branches	468 units	$5.00/year	2,340
Reorder cost			
Branch	52/year	$19.00	990
Factory	26/year	$13.50	350
Total			$5,780

weeks, but experience showed that in 60% of the weeks no orders were placed, in 30% one order, and in 10% two, three, or more orders were placed. EXHIBIT VIII shows orders on the factory and the production level for a representative period of weeks.

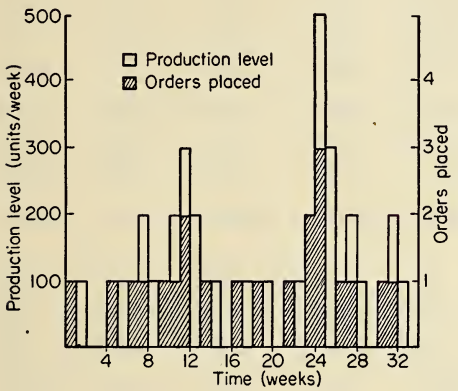


EXHIBIT VIII

FACTORY ORDERS AND PRODUCTION LEVEL

Factory snarls due to these fluctuations occasionally caused the factory to miss deadlines. These in turn led on occasion to warehouse delays in filling branch orders, and forced the branches to hold to the two-week delivery time even though actual transit time was only one week. An analysis revealed the following:

Factory fluctuations were very costly. A statistical regression of costs against operating levels and changes showed that annual production costs were affected more by the average *size* of changes in level than by the frequency of change; a few large changes in operating level were much more costly than many small changes.

Under the "economical reorder quantity" system, production fluctuations were no larger than before, but the average change up or down actually equaled 80% of the average production level. This was estimated to cost $11,500 annually, bringing the total cost of the system, including costs of holding inventories, placing orders, and changing production rates, to $17,280 per year.

This led to the suggestion that the company try a new scheme so that orders on the factory warehouse and the factory would be more regular. A system with a fixed reorder cycle or period was devised, under which branch warehouses would place orders at fixed intervals, the order being for the amount sold in the period just ended. The factory warehouse would ship the replenishment supply, order an equivalent amount from the factory, and receive the order within two weeks or by the beginning of the next review period, whichever was longer.

Under this scheme, each branch warehouse would need to keep its stock on hand or on order sufficient to fill maximum reasonable demand during one review period plus delivery time (tentatively taken as two weeks) on the basis of the reorder rule described previously in this article. The question to be determined was: How long should the review period, that is, the time between reorders, be? EXHIBIT IX summarizes inventories and costs for reorder intervals ranging from one to six weeks, based on the following facts and figures:

(1) *Branch safety stock* was determined from a study of branch sales fluctuations, to allow for maximum reasonable

EXHIBIT IX

SUMMARY OF REORDER PERIOD COST COMPARISONS

	Length of period (weeks)					
	1	*2*	*3*	*4*	*5*	*6*
Branch warehouse						
Safety stock	24.0	26.0	27.0	28.0	30.0	31.0
Cycle stock	12.5	25.0	37.5	50.0	62.5	75.0
Transit stock	25.0	25.0	25.0	25.0	25.0	25.0
Total units of stock	61.5	76.0	89.5	103.0	117.5	131.0
Annual inventory cost	$310	$380	$450	$515	$590	$650
Ordering cost	990	495	330	250	195	165
Total cost each branch	$1,300	$875	$780	$765	$785	$815
Total cost four branches	$5,200	$3,500	$3,120	$3,060	$3,140	$3,260
Factory warehouse						
Safety stock	33	33	41	47	52	58
Cycle stock	50	100	150	200	250	300
Total units of stock	83	133	191	247	302	358
Annual inventory cost	$290	$465	$670	$865	$1,060	$1,250
Ordering cost	700	350	235	175	140	120
Total cost factory	$990	$815	$905	$1,040	$1,200	$1,370
Production change costs	$1,600	$2,250	$2,760	$3,180	$3,560	$3,900
Total system costs	$7,790	$6,565	$6,785	$7,280	$7,900	$8,530

demand over the reorder interval plus the two-week delivery period.

"Maximum reasonable demand" was defined to allow a 0.25% risk of being out of stock in any one week (equal to the 1% risk on the average four-week interval under the "economical reorder quantity" system described previously).

(2) *Branch cycle stock* would average one-half of an average shipment. Under this system, the average shipment to a branch each period would equal the average sales by the branch in one period (25 units × number of weeks).

(3) *Transit stock* equaled one week's sales.

(4) *Branch inventory carrying cost* was $5 per unit per year.

(5) *Branch ordering costs* equaled $19 per order, with one order per period. A one-week period would mean 52 orders per year; a two-week period, 26 orders per year; etc.

(6) *Factory safety stock* was set to allow a 1% risk that the warehouse would be unable to replenish all branch shipments immediately.

(7) *Factory cycle stock* in process or in the warehouse would be approximately equal to one-half the sales in any one period.

(8) *Factory inventory carrying cost* was $3.50 per unit per year.

(9) *Factory ordering costs* equaled $13.50 per order (see 5 above).

(10) *Production change costs* were proportional to the period-to-period changes in production level, equal under this sys-

tem to period-to-period changes in branch sales.

The figures show that a two-week reorder interval would be most economical for the company as a whole, and this was chosen. Costs were estimated to be $6,600, compared with $17,300 under the "economical reorder quantity" system. While the new system cut total inventories by nearly 70%, most of the gain came from smoother production operations. Exhibit x shows weekly production for a representative period under the new system.

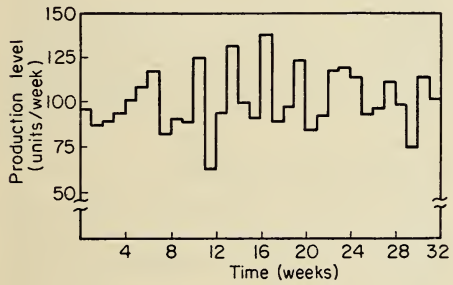


EXHIBIT X

PRODUCTION FLUCTUATIONS REDUCED WITH FIXED REORDER CYCLE

Further economies became apparent when the system was in operation:

(1) The reduction in production fluctuations made it possible to meet production deadlines regularly, cutting the effective lead time in deliveries to branches and thereby permitting modest reductions in branch safety stocks.

(2) The inventory system was found well suited to "open" production orders. Instead of issuing a new order with each run, the moderate fluctuations made it possible to replace production orders with simplified "adjusting memos" and at the same time to eliminate much of the machine setups.

"BASE STOCK" SYSTEM

The success with the periodic reordering system encouraged the company to go further and try the "base stock" system referred to earlier. Under this system, the branch warehouses would *report* sales periodically. The factory would consolidate these and put an equivalent amount into production. Stocks at any branch would be replenished whenever reported sales totaled an economical shipping quantity.

Two possible advantages of this system compared to the fixed period scheme were: (1) Branches might be able to justify weekly sales reports, reducing production fluctuations and safety stock needs still further. (2) It might be possible to make less frequent shipments from factory to branches and make further savings. The following questions had to be decided:

How frequently should branches report sales? As noted earlier, cost studies showed that of the $19 total cost of ordering and receiving goods $6 represented clerical costs in placing and recording the order. Here is a summary of the costs affected by the choice of reporting interval:

	Reporting Interval			
	One Week		*Two Weeks*	
	Number	*Cost*	*Number*	*Cost*
Branch safety stock	100	$ 500	108	$ 540
Production changes		1,600		2,250
Branch clerical costs	4 × 52	1,250	4 × 26	625
Total		$3,350		$3,415

Thus, there appeared to be some advantage to reporting sales weekly from branches to the factory.

How big should replenishment shipments be? Exhibit xi summarizes the system costs related to the size of shipment from factory to branch. Each line shows the total of the cost indicated plus those represented by the line below. The total system cost (top line) is lowest at 82; that point is therefore the optimum shipping quantity from factory to branch warehouse. The same answer can be obtained from the formula given before, $\sqrt{2 \cdot \$13 \cdot 1{,}300/\$5} = 82$.

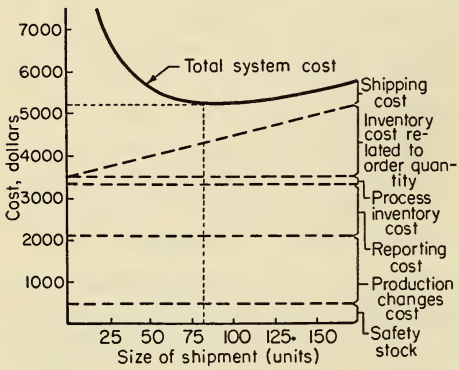


EXHIBIT XI

OPTIMUM SHIPPING QUANTITY FROM FACTORY TO BRANCH WAREHOUSE UNDER BASE STOCK SYSTEM

The base stock system therefore was set up with weekly reporting and replenishment shipments of 82 units to branches. The total cost of the base stock system was $5,200 compared with $6,600 under the previous system.

STABILIZED FURTHER

The company, cheered by its successes, decided to see if even further improvements might be obtained by cutting down further on production fluctuations. As it was, the production level under the base stock system was being adjusted each week to account for the full excess or deficiency in inventory due to sales fluctuations. It was proposed that production be adjusted to take up only a fraction of the difference between actual and desired stocks, with added inventories used to make up the difference.

The possibilities were analyzed along the lines described previously in the text; the results are summarized in Exhibit xii. The two costs that would be affected are costs of changing production and costs of holding inventories in particular safety stocks. These are affected by the fraction of the inventory departure that is made up each week by adjusting production.

The study showed that the cost would be minimized with the rate of response set equal to 0.125, as seen in

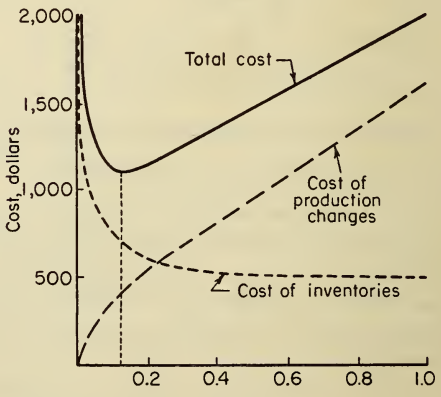


EXHIBIT XII

COST OF PRODUCTION CHANGES AND SAFETY STOCK VS. RATE OF RESPONSE TO SALES FLUCTUATIONS

the exhibit. (This compared with a response rate of 1.0 under the base stock system.) The additional savings of $970 brought the annual cost of the system down to $4,200.

SUMMARY

The results of all the changes made by the division management were substantial:

(1) *A major reduction in stocks*—They had been cut 35% from what they were even with the "economical reorder quantity" system.

(2) *A substantial reduction in production fluctuations*—Exhibit xiii shows what weekly production levels for a typical period looked like at the end, contrasted with Exhibits viii and x for the same sales.

The problems of the case are common even among the best-run businesses and can be solved in much the same way with much the same results. Of course, a large part of the effort and expense that were necessary in this step-by-step, evolutionary approach could be saved. Technical methods are available for analyzing and measuring the performance of alternate systems so that management can proceed directly to the ultimate system that is most desirable; management does not have to feel its way. Let me emphasize again, however, that no one kind of system should be considered "the goal." The efficiency of any given inventory control plan depends too much on the demand and cost characteristics of the business. . . .

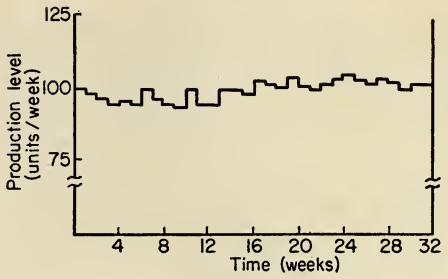


EXHIBIT XIII

PRODUCTION LEVEL UNDER THE BASE STOCK SYSTEM WITH A REACTION RATE OF 0.125

✦✦✦✦✦✦✦✦✦✦✦✦✦ PRODUCTION SCHEDULING: *an operations research case study*

RUSSELL L. ACKOFF

The case study I am going to present is primarily directed toward emphasizing two aspects of Operations Research. The first aspect is the *breadth* of the OR approach to problems, its attempt to consider the effect of a policy decision on all phases of an organization's operations; in other words, the inclusion of the widest possible range of variables. The second aspect involves the concept of *continuous* or *dynamic* research: that is, the

Advanced Management, *March 1955, XX:3, 21–28.*

continuous extension of the scope of the research as it proceeds, generating new problems with the tentative solution of each previous one. Thus, problems are never *solved* in any absolute sense; the research is designed to enable the organization to proceed from one level of operating efficiency to a higher level in successive and progressive steps.

This particular case began in June 1952, when the Operations Research Group at Case Institute was invited by the President of the Warner and Swasey Company to speak with him and several other executives.

First let me tell you a little bit about the Warner-Swasey company. It is the world's largest producer of turret lathes, turning out better than 50 percent of the total of this country's output. The company also produces several other lines, some related and others unrelated to its major product. Turret lathes vary in price from $10,000 to $40,000. In 1952–53 the company did a total business in the neighborhood of $55,000,000. Early in 1954 employment in its plants was about 3500. Employment was higher during the war, but even now the plants are operating with several shifts.

The executives of the company had no one particular problem in mind. We suggested that before selection of a problem was made we should know much more than we did (we knew practically nothing) about the company's operations. Consequently, we suggested that a few weeks be devoted to orientation by our group—that we spend this time getting familiar with the company and formulating possible OR problems. This suggestion was accepted by the executives with no hesitation.

A team of three was established. It included two from Case Institute's OR Group and one member of the company. I would like to say categorically that any success the group may have had is due primarily to the contribution of the company member and the cooperation of company personnel at every level of its operations. The company member of the OR team was experienced in corporate financial research. He is a sort of trouble-shooter on the staff of the company's treasurer, and as such has dealt with a variety of messy problems involving every phase of the company's operations. Though he had had no previous contact with OR, his wide experience and preoccupation with *methods* of problem-solving made him an ideal member of the team.

We were given space in the treasurer's office though we continued to report directly to the president of the company. The treasurer, however, greased many skids for us. The team has varied in size. It has included as many as four professionals from Case, graduate student assistants, and varying numbers from the company. Throughout, consultation has been frequent with all other members of the Case OR group, other members of the faculty, and a wide variety of personnel from the company.

The initial job of at least the Case members of the team was one of orientation. First a comprehensive tour of the main plant and administrative offices was arranged. Then we asked to

see organization charts. The company is not "chart-happy" and consequently there was some difficulty in obtaining a chart. Once it was obtained, our questions soon demonstrated how unrevealing such a chart can be regarding operations. Consequently, we asked ourselves what it was we wanted to know about the company. We wanted to know (1) the *nature of the operations* in which they are involved, and (2) how they *control* these operations.

Now, as soon as we started to toss around the term "control," the approach of cybernetics—the science of control through communication—came to mind. In effect we decided to consider the organization as a communications circuit which controls a productive process. Where, then, is the ultimate source of information that flows through the circuit? It is the customer, the user of the product. How is information transmitted from customer to company? Through sales engineers.

We began our orientation, then, in the sales department. We learned how salesmen selected potential customers, what type of contact they made with customers, how they reported their activities, how orders were prepared and so on. Then we examined the processing of this information through the various sections of the sales department, and learned how the processed information was put into the production system. We saw how the information eventually came to pick up raw material, transform it, and eventually yield a product that was shipped to the customer.

At the end of two weeks we had reams of data and forms. This had to be digested. We spent several days extracting the essence of this complex process and recording it in a "Control and Materials Flow Chart." (See Figure 1)

It would take too long to explain this chart in detail, but I should like to explain one part of the circuit illustratively. The production planning department receives an assembly schedule each month. This schedule shows the numbers and types of units to be shipped for the next five months. For each type of unit scheduled, the production planning department has a complete list of required parts. Further, for each part this department has a file card which shows how many are in stock, in production, or on order from vendors. For any one part there are four possible situations which can exist: (1) it is produced by the company and is in stock, (2) it is produced by the company and is not in stock, (3) it is purchased and is in stock, and (4) it is purchased and is not in stock. Let us only consider here what happens in the fourth case.

Production Planning prepares a list of all the out-of-stock purchased parts for a given model of turret lathe. This list is called a "Traveling Requisition" and is sent by the Production Planning Department to the Purchasing Department. The Purchasing Department prepares seven copies of orders for each part and returns the Traveling Requisition to the Production Planning Department as a notice that the orders have been placed. The original copy of the order goes to the supplier. One copy goes to the Cost Analysis Department which eventually uses this in-

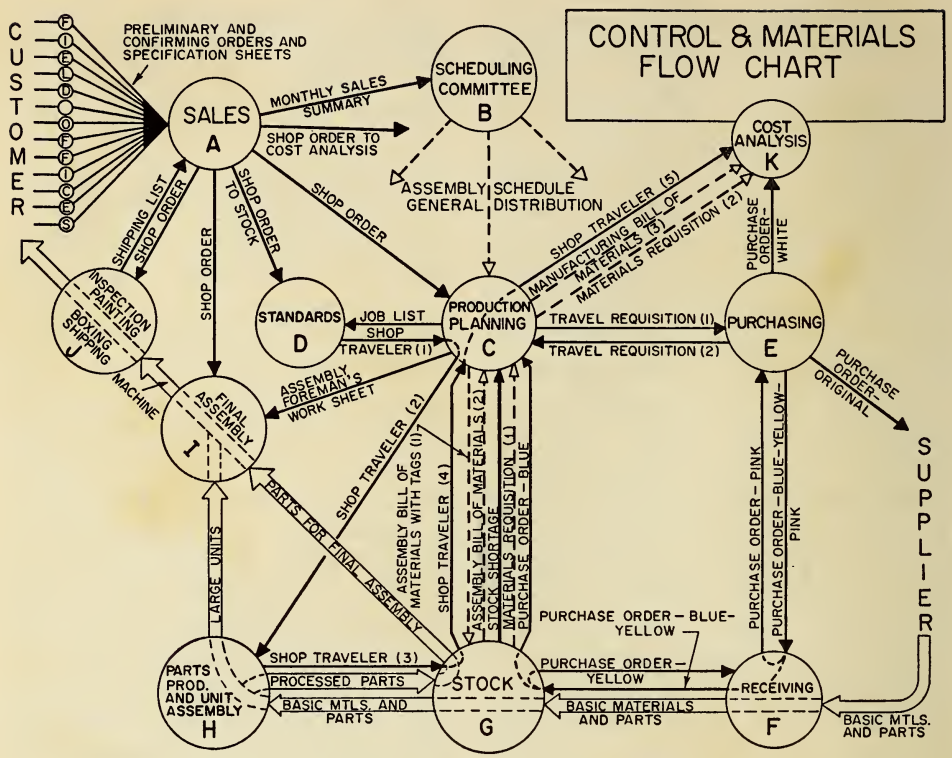


FIGURE 1

formation with other information to determine unit production costs. Three copies are sent to the Receiving Department. The remaining two copies are placed in a moving file in the Purchasing Department to enable them to keep their fingers on delayed deliveries and follow them up.

When the Receiving Department gets the ordered parts from the supplier, it returns one copy of the order to the Purchasing Department. The Purchasing Department puts the order into an inactive file. Its job is completed. The Receiving Department sends the parts along with its two remaining copies of the order to the Stock Room. Stock Room personnel receipt one copy and return it to Receiving, whose job is now completed and recorded. When the parts are entered in stock the last copy of the order is sent by the Stock Room to the Production Planning Department. This department notes the availability of the part on that part's stock card. This circuit is now complete.

The circuits for each of the other parts-possibilities are also shown on the chart, along with other phases of the process. Not only did we find this chart useful, but the company's executives have found it fruitful in discussions of

organization, and for orienting new employees and visitors. Consequently, several large copies were made and are in use by the company.

PROBLEM OF INVENTORY LEVELS AFFECTS ALL DEPARTMENTS

In the process of collecting the information necessary for preparing the analysis the chart represents, the team began to get a "feel" of a problem that concerned just about every department. This was not surprising since the problem involved inventory levels.

As might be expected in connection with a product that ordinarily has a highly variable demand pattern there was considerable concern with the risks involved in carrying a large inventory. In a period when high volume requires a high level of inventory, concern with the costs of carrying too low an inventory is shoved into the background and active planning is concentrated on the first aspect of the problem.

The team obtained records of the physical inventory taken at the end of 1951 and analyzed them by product and inventory class. The analysis made certain obvious things more obvious; for example, that 65 per cent of the inventory was devoted to turret lathes. The study also disclosed a not-so-obvious fact: 29 per cent of the inventory was invested in parts in process and finished parts for turret lathes. On the basis of these facts, and the fact that an inventory problem seemed to be a good way to get into company operations, we decided to recommend a study of the turret-lathe-parts inventory.

INVENTORY LEVEL AFFECTS PRODUCTION COSTS NOT SALES VOLUME

We met again with the company's executives, showed and discussed the Control and Materials Flow Chart, and suggested the parts-inventory problem. They agreed with the suggestion and we were "turned loose."

We probed to find out what people in the company took the parts-inventory problem to be. The then current formulation was: What is the minimum parts-inventory necessary to maintain our present level of shipments?

The OR team suspected this formulation of the problem because it assumes that the margin of profit on sales is constant or, at any rate, if it varies, its variations do not depend closely on the inventory level. If, as we suspected, the size of the inventory can be used to reduce production costs with direct results on operating profits, it seems that the size of inventory should be determined not as the least amount necessary to support a given volume of sales, but as the amount which can be used to yield the greatest profit at the given sales volume. Such reasoning led us to reformulate the problem as one of developing a method of scheduling the production of parts in such a way as to yield greatest profits.

What is involved in the production of a part? Here our orientation showed its worth. First there are the raw materials whose values are composed of purchase price plus freight costs. Then there is a raw material inventory stage in which more money is invested in the materials. Then there is a planning

stage in which the future of the material is determined. This planning also involves a cost. The shop must be set up for producing the part. The material must be worked on, and it must wait between operations. Then there is a finished parts inventory.

On the basis of a preliminary study we decided that raw material and in-process inventory would be little affected by changes in the production system. To simplify the problem, we assumed this to be so. Subsequently we came back to these assumptions. But more on this later.

This loose description of the production of a part had to be tightened. Such tightening was brought about by studying the current scheduling of parts and by identifying and defining the pertinent variables in the process. The parts were scheduled monthly in this company (i.e., a one-month scheduling period). It was convenient to have some time interval relative to which costs were computed. The period of one-year was selected and was referred to as the *planning period*. The model developed is general in the sense that the scheduling and planning periods can be set at any specified interval.

The following variables were identified:

C_1 = set-up and take-down cost per lot.
C_2 = raw material cost plus process cost per part.
P = inventory carrying cost expressed as a per cent per month of the value of the part.
L = required number of parts per scheduling period.
m = number of scheduling periods.
$N = mL$ = number of parts required per planning period.
$N' = mL/n$ = number of parts per lot.
K = total incremental production cost per planning period (i.e., total cost less raw material inventory cost and in-process inventory cost).
K' = total incremental production cost per lot.
n = the number of lots per planning period.

The meaning of at least some of these variables is far from obvious, so let us consider them one at a time.

A DEFINITION OF VARIABLES THAT MAKE UP THE PROBLEM

First let us consider C_1, the set-up and take-down cost per lot. The term "lot" refers to all the parts which are made for a single set-up of the machines. The size of a lot may vary. In other words, it is the number of parts scheduled for a given set of operations. The set-up and take-down cost includes a number of components, four major ones: (1) Office set-up: before anything is done in the shop, the Production Planning Department must schedule that production and the standards Department must prepare necessary drawings and control forms. This preparation costs money. It took some study to isolate and measure this cost, which is independent of the number of parts scheduled, because it is a paper operation. (2) Shop set-up cost: this consists of the cost of actually adjusting

the machines to perform the needed operations, the cost of the scrap which is involved in making adjustments at the beginning of the run, and the cost of setting up the quality inspection procedure. (3) Shop take-down costs: this involves the cost of entering the finished parts into stock and performing the necessary paper work attached thereto. (4) Office take-down: this is the cost of the analysis performed by the cost analysis section, a process which involves the use of I.B.M. equipment.

It is apparent that the job of estimating the value of the variable, C_1, for any specific part is not simple. It required a good deal of work with a number of departments. This work had a good effect, however, for it raised an important question. The cost accounting system did not lend itself easily to providing values of this variable. Shouldn't it be equipped to do so? The company's new comptroller used this question to reinforce his effort to convert the accounting process from one which in the main presented passive historical statistics to one which provided active control-data. The need for functional accounting was supported by our efforts. Subsequently we were able to assist in the conversion in a small way.

The second cost listed, C_2, is also a combination of two costs which are ordinarily treated separately. It became convenient at this point of our study to group them. The first component is the cost of the raw material used in making the part. The second is the process cost; that is, the cost of direct labor expended in working on the material, plus overhead. Overhead costs, which are included in C_1 as well as C_2, were not easy to determine. A satisfactory preliminary estimate was obtained which expressed this cost as a function of man-hours of direct labor involved in the operation.

The third cost is that of carrying goods in inventory. We went through the literature in an attempt to find a way to estimate this cost and we consulted with an expert accountant. But all this effort was in vain. We were forced into doing the job for ourselves. It was done as follows:

DETERMINING COST OF CARRYING GOODS IN INVENTORY

A study was made of the cost involved in running one of the company's warehouses. We took account of rent, heat and light, alarm service, wages, supervision, supplies, and depreciation. The ratio of the sum of these costs per month to the value of the parts stored was 0.88 per cent. To this was added the cost of borrowing the capital invested in the inventory. This yielded a figure slightly more than 1 per cent per month per dollar invested in stock. For safety's sake in subsequent analysis a pessimistic figure of 2 per cent and an optimistic figure of 0.5 per cent were also used. The effect of so doing will be considered later.

The next variable in the list, L, is the required number of parts per scheduling period. In this company the scheduling period is one month; that is, a new assembly schedule is issued each month. Therefore, withdrawal of parts from stock for assembly occurs once a

month. That is, withdrawals for assembly are discontinuous and occur once each month.

Since the company is working against a backlog of orders, the monthly requirements will remain relatively fixed until that backlog has disappeared. This simplifies the initial problem, but subsequently (as later discussion will show) the handling of variable demand was taken into account.

FIND THE NUMBER OF LOTS THAT MINIMIZES PRODUCTION COSTS

The next variable, m, represents the number of scheduling periods per planning period. We took one year to be a planning period. It turns out, however, that the scheduling procedure eventually derived is independent of this variable.

The remaining variables are more or less self-explanatory. N, the number of parts required per planning period, is equal to the product of the number of scheduling periods per planning period and the requirement per scheduling period; that is mL. N', the number of parts per lot, is equal to the requirements per planning period (mL) divided by the number of lots per planning period (n); therefore, $N' = mL/n$. The total production cost per planning period and per lot (K and K′) need no explanation.

The variable n, the number of lots per planning period, is critical because it is the "manipulation" variable; that is, it is the aspect of the system which can be set so as to yield varying costs. The problem is to find the value of n for which the total annual production cost is minimum.

Once the variables are identified the problem is to relate them. We can begin by considering the total cost per lot. This cost can be broken down into the sum of four components:

(1) Set-up and take-down cost per lot.
(2) Raw material and process cost per lot.
(3) Inventory cost on the investment in set-up and take-down cost per lot.
(4) Inventory cost on the investment in raw material and process cost per lot.

Each of these components can be translated into the symbols we have introduced. The details of the translation need not concern us here. The result, however, is the following equation:

$$K = nc_1 - mLc_2n - \frac{nPc_1}{2}\left(\frac{m}{n} - 1\right) - \frac{Pc_2Lm}{2}\left(\frac{m}{n} - 1\right). \qquad (1)$$

K is the total incremental production cost per planning period.

Now, for any part, the values of c_1 (set-up—take-down costs), c_2 (raw material plus process costs), P (inventory rate), L (monthly parts requirement), and m (number of scheduling periods per planning period) can be determined. The problem is to find the optimum value of n, the number of equally sized lots to be run per year, when the above quantities are known. This can be determined graphically. That is, the total annual cost for a number of values of n could be computed and plotted on a graph. The results

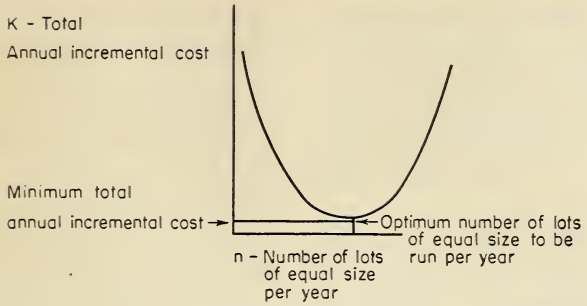


FIGURE 2

would look like the curve shown in Figure 2.

The optimum value of n (n_0) is one for which the total annual cost is minimum. The graphic solution does not give the exact value of n required: the methods of differential calculus enables us to derive an equation which yields the exact optimum value of n. We need not be concerned with the mathematics here as long as we realize that we are simply finding the value of n that minimizes the value of K. The resulting equation for n_0 is as follows:

$$n_o = m \sqrt{\frac{LPc_2}{c_1(2-P)}} \qquad (2)$$

The value of n_0 given in the equation above is the most economic number of equally sized lots to be run per year. To obtain the optimum lot size, N'_0, an algebraic translation yields the following equation:

$$N'_o = L/\sqrt{\frac{LPc_2}{c_1(2-P)}}$$

Note that the optimum lot size is independent of M, the number of scheduling periods per planning period, and hence is independent of the length of the planning period.

Equation (1) for total incremental cost per planning period has a limitation on its usefulness: it is exact only if an integral number of months' requirements are made per lot run. If, for example, 1½ months' requirements are included in a lot, this equation only gives an approximate value of the total incremental cost. Furthermore, it is not practical to schedule non-integral multiples of monthly requirements per lot. This practical difficulty can be shown in this way. See Figure 3.

If, for example, 1½ monthly requirements are made, 1 month's requirement

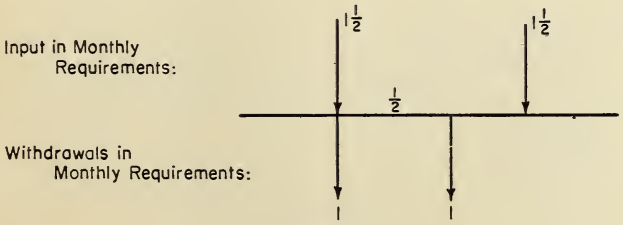


FIGURE 3

is withdrawn almost at once, leaving 1/2 month's requirement in inventory for a month. But at the end of the month another month's requirement is needed, and only 1/2 is available. This suggests the modification in scheduling shown in Figure 4.

That is, by scheduling 1½ monthly requirements for two consecutive months, and skipping a month, the difficulty which arises from the procedure shown in Figure 3 is overcome.

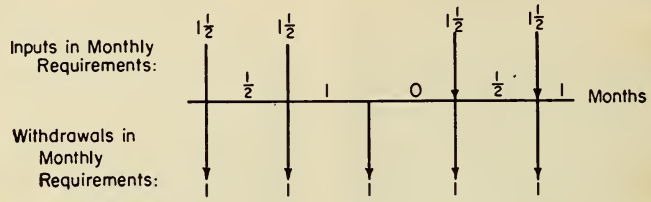


FIGURE 4

But it becomes apparent that the 1/2 monthly requirement inventory carried the first, fourth, etc., months can be eliminated if the procedure shown in Figure 5 is used.

The procedure shown in Figure 4 is the most practical. It can be generalized, and suitable modifications in equations (1) and (2) can be made.

The first three terms of the resultant revised equation (1) are the same as these shown in equation (1). Total set-up costs, material and cycle costs, and inventory on set-ups remain unchanged. But inventory on material and cycle costs are increased. This increase, for the problem at hand, turned out to be negligible, consequently we could use equation (1) to compute costs, since it is simpler to use; but we could schedule in integral lots according to the procedure just described.

The economic lot size equations considered above represent no unique contribution of OR. Industrial Engineers have been using such equations since about 1920. What is unique to the OR approach is yet to be presented; it is what comes after the derivation of these equations. More explicitly, it is the generalization of the approach to handle variable demand and to include costs associated with numerous other phases of the company's operations.

When these optimum economic lot size equations had been completed, the OR team met once again with the executives of the company. The mathe-

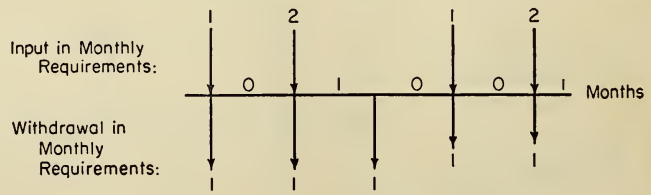


FIGURE 5

matics was not discussed but the underlying ideas were. The meeting brought out a good deal with respect to the definition of costs, and ways of obtaining estimates of the various costs. The executives decided that it would be worth trying out the equations. The Production Planning Department agreed to select twenty-three parts for this purpose. No systematic sampling went into their selection; rather, the parts were selected because they presented a wide variety of scheduling problems.

After the parts were selected, the team computed the total annual cost of, and set-up time for producing each, using then current scheduling practices, and also computed total annual cost and set-up time assuming production in most economic lot sizes. The results indicated a large potential saving in both costs and time. To obtain these savings, it was indicated that inventory practically had to be doubled. That is, by increasing the finished parts inventory to about twice its then current size, substantial savings in time and money were indicated.

Another meeting with the executives brought agreement that the results obtained were of such a nature as to merit further study. We decided to study intensively a sub-assembly unit consisting of 112 parts, and make a comparison similar to the one made for the 23 parts. In this study a comparison was first made between scheduling one month's requirements per month and the optimum procedure. The results indicated a reduction in production costs of approximately 10 per cent, and a reduction in set-up time of approximately 85 per cent. Of this potential reduction of 10 per cent in production costs, then current scheduling procedures were obtaining 6.5 per cent, leaving a further potential reduction of 3.5 per cent.

The Scheduling Department had been using a lot of the size required for one month as "normal" and departing from this normal wherever it appeared economies in purchasing or processing would result from larger lots. This meant they were bucking the psychological hazard of deciding to increase inventory when they scheduled larger lots. The economic lot calculations were closer to their actual pattern of scheduling than the "normal" they were used to. Since the economic lot standard was higher than actual practice it put them in the psychological position of keeping down the company's money "tied up in inventory." In effect, use of the economic lot as a standard made it necessary to justify reductions of inventory from the new standard rather than increases from the old one.

Management considered the results of this comparison significant enough to warrant institution of an experimental program for conversion of production to optimum lot sizes.

To some this might seem like the end of the role of Operations Research. But in fact it was in a very real sense "only the beginning." The most difficult and characteristically OR aspects of the problem arose once implementation was accepted as something to be desired. All the succeeding problems cannot possibly be covered in detail, but their nature can be indicated, and in some

cases how they were or are being handled.

First, we assumed in the development of the parts production model that the in-process inventory was not significantly affected by changes in production scheduling. This assumption had to be examined. An equation expressing the cost of carrying in-process inventory was developed. By the use of this equation, we determined the increase in in-process inventory cost produced by optimum scheduling compared to current practices. It turned out to be negligible, thereby justifying the initial assumption.

Secondly, an increase in inventory brought about by moving towards more economic lot sizes requires additional capital. This raises three questions: (a) How much money is needed? (b) At what interest rate can it be obtained? (c) How would this increase in borrowed capital affect the credit and financial standing of the company? A study was conducted to answer these questions.

OPTIMUM SCHEDULING CAUSES PRODUCTION BOTTLENECK

Next, we assumed that the cost of raw material would not be affected by changing the production schedule of parts. The results obtained indicated the possibility of ordering raw materials in larger quantities and benefiting thereby from cost reductions. Investigation showed this would only be true for a small percentage of the parts. But we learned, in this inquiry, that freight costs could be significantly reduced on certain parts. Trucks are used to haul some parts. For example, the cost for hauling forty parts in some cases is virtually the same as for hauling ten parts. It was not considered practical to take resulting changes in freight costs into account in calculating economic lot sizes, it was considered that this was one more benefit to be expected as economic-lot-size production is approached.

A whole group of problems involving the mechanics of scheduling arose. For example, we found out that one part should be scheduled so that eight months' requirements should be produced at a time. This part requires almost all of the work done on it to be done in one shop section. If eight months' requirements were put through this section at a time, no other parts requiring work by this section could be processed for about a month. This situation represents what the Operations Researcher calls a queueing or waiting line problem. Unfortunately, in this case, the situation was much too complex to apply available techniques for handling such problems. To get an idea of the complexity of this in-process waiting line problem we constructed on paper a small model company, one very similar except in size to the company involved. We scheduled through several years to make a conversion to optimum scheduling. Though only four parts were involved in this paper operation, this dry run enabled us to anticipate most scheduling problems involved in conversion in the Warner and Swasey Company. This experience provided an insight into some of the factors which in practice would require

less than economic lot size scheduling.

I shall mention only briefly some of the practical scheduling problems. Though optimum scheduling, when in operation, would require less than current production hours for the same output, it requires more to get over the "conversion hump" because larger runs for some parts must be started while others continue to be produced by the current procedure. The conversion can be accomplished gradually by using what "play" is available in the current scheduling, or by capital expansion, or by sub-contracting parts until the hump is mounted.

Economic lot size computations were very valuable in this study because it provided a basis for computing the reductions in production cost that increased facilities would yield. This reduction was compared with cost of the new facilities.

The policy of gradual conversion was selected. It was estimated that three years would be required to complete the conversion by this method. Possible effects of increasing and decreasing volume of business on conversion were studied and plans for meeting such contingencies were developed.

CHANGING ORDER OF PROCESSING REPAIRS GIVES MORE SPACE

A study of the storage facilities available for the parts in this model is necessary to determine for which parts additional storage space is required, and how much. It also provides a basis for determining which parts can now be produced in larger runs without creating a storage problem. Concern with additional storage space turned our attention to the process of filling orders for repair parts. A team set up to study this problem found that by modifying this repair order processing, storage space currently used in that process could gradually be made available for storage of finished parts. This study also showed how to alleviate shortage problems arising out of borrowing parts for assembly to fill repair orders.

The clerks who do the actual scheduling are not equipped to handle mathematical formulae. Consequently, at first graphic devices (nomographs) were developed which would permit them to determine optimum run sizes very simply. But even with a nomograph, the clerk needs basic cost and requirement figures. Arrangement with the Cost Analysis and Sales Departments are necessary to get current data. Arrangements for recording the data in a convenient way are also necessary. The company is in the process of converting to the use of automatic equipment for scheduling. Studies have indicated that the mathematics involved in optimum scheduling can be handled accurately, quickly, and economically on this equipment. The nomograph is still useful, however, for checking purposes and for handling certain problem-parts.

In arranging for the supply of required cost data it became increasingly apparent that costs are subject to variation due to short range changes in the production process and to the accounting interpretation of these changes.

The extent of such variation had to be determined along with its possible

effect on the new scheduling procedure. A study of this variation was made. It showed that within the range of costs to be expected it was still profitable to schedule in the way described. For example, it was shown that inventory cost (P) would have to be greater than 3 per cent per month, on the average, for then current procedures to be better than those described here as optimum. It will be recalled that an inventory cost of 2 per cent per month was considered to be a maximum possible inventory carrying cost (on the basis of an earlier study).

SALES FORECASTING STUDIED AS MARKET CHANGE INDICATOR

The model and procedure described was geared to the company's then current requirements, which were relatively constant. Even now the demand is becoming variable. Can the procedure be modified so as to handle variable demand? It can be, providing a reliable prediction of sales is available and providing the precision of such estimates is known. For this reason we made an extensive study of sales forecasting. A number of methods using internal and external variables (published indexes) were subjected to comparative study. Our results were, I think, very complimentary to the company. An adjusted estimate based on the forecasts provided by the company's sales offices yielded best results in the sense that such an estimate is unbiased and has more reliability (less variance) than any of the many other methods tested. Furthermore, the errors of the estimates turned out to be normally distributed. This makes the estimates convenient to handle mathematically in scheduling equations.

Not only were we concerned with short range month-to-month forecasting for scheduling purposes, but we were also concerned with longer range forecasting for planning purposes, in particular, with forecasting changes in market trends. We were very surprised when we found a very sensitive indicator of changes in market. It consisted of applying a statistical quality control method to the market. Twelve actual net sales figures for each 90-day period beginning with the first day of each month of a year are plotted graphically. A trend line is fitted to these points by the method of least squares, and is projected ahead for another year. The standard deviation of the twelve points around the line are computed. Then two lines are drawn parallel to the trend line on both sides of it at a distance of two standard deviations from that trend line. Less technically, a band is drawn about the trend line such that if market variations were random one would expect approximately 95 per cent of the actual sales figures to fall within the band. Now as sales figures come in for each new 90-day period they are plotted. If one of these points goes above the band an improved trend is predicted, if below, a worse trend is predicted. This method was applied retrospectively over a number of years and not a single false prediction was noted. The method has been applied prospectively as well as retrospectively and has already correctly picked up a change in market trend.

There are other indications besides

a point going outside the band which are useful. For example, if actual sales are a random variable, it is very unlikely that five consecutive figures will fall above or below the trend even within the band. Such an occurrence turns out to be a good indicator that a figure will shortly go outside the band.

A variation of the method just described was also developed which predicts changes in market trend with equal reliability, but on the average does so better than two months earlier. In this refinement, once the trend line is determined for the last twelve 90-day figures, the corresponding adjusted 90-day predictions prepared by the sales office were plotted about this line. Then the standard deviation of the predictions about the trend line is computed. The band limits are then drawn 3 of these standard deviations away on either side of the trend line. Subsequent predictions are then plotted, and the results are analyzed in the same way as is done in the first method. This method of predicting changes in market trends has since been successfully applied to another entirely different business.

OPERATIONS RESEARCH—A BROAD APPROACH TO ALL BUSINESS PHASES

Approval to begin conversion to economic lot sizes on one model of the turret lathe was given several months ago. The conversion is a difficult one involving many detailed problems and it is not being conducted in a vacuum. Other changes resulting from decisions made in the engineering and sales departments and from the regular changes in the shop are taking place at the same time. Even after conversion to the new method of scheduling is completed the exact computation of savings resulting will not be possible. This is especially true because one of the resulting benefits to be expected is an increase in control over the process which will permit of better coordination with the other types of change which accompany it in time.

While the conversion is going on the OR team is attempting to generalize the scheduling procedure to include production of repair parts. Here no forecasts are available and demand is extremely variable. Consequently, a fresh approach is required in this area.

I hope this report has given you some feeling for Operations Research and has impressed on you the breadth of its approach, its successive inclusion of more and more phases of an organization's operations in the study of any phase of these operations. It is this breadth which makes Operations Research such a useful management tool.

B

Operations Research in Marketing Management

DETERMINING OPTIMUM ALLOCATION OF SALES EFFORT

RUSSELL L. ACKOFF

This report covers an Operations Research project conducted jointly by General Electric Lamp Division personnel and members of the Operations Research Group at Case Institute of Technology. The project was concerned with determining how sales time could be used by the Division so as to provide greater returns on investment in this valuable human resource. . . .

INITIAL PLAN OF THE RESEARCH

Before salesmen can operate most efficiently, three questions must be answered:

(1) How frequently should a salesman call on any specific account?

(2) How many accounts should be assigned to a salesman?

(3) What kind of man makes the best salesman for what type of account?

Proceedings of the Operations Research Conference, Society for Advancement of Management, Advanced Management, *1955, 74–85.*

Obviously these three questions are not independent. Any policy regarding the use of salesmen assumes either implicitly or explicitly some kind of answer to all three. In the interest of producing usable results in as short a time as possible, it seemed reasonable to concentrate attention on the first question. This, then, was the specified objective of the project: to accept the present sales force and their territorial assignments and determine how the average salesman should allocate his time among his accounts so as to maximize the sales volume. Information pertinent to the other two questions turned up incidentally in pursuit of an answer to the first.

The research plan began with the assumption that, in general, sales volume increases with an increase in the sales time. But it did not seem reasonable to assume that sales volume increased in direct proportion with increases in sales time. It seemed more reasonable to assume the type of relationship which can be represented graphically by an "S-shaped" or "learning" curve. (See Figure 1.) The slope of this curve is of particular importance since it represents the rate of increase in sales volume related to increases in sales time. Also important is the point at which the curve begins to flatten out and reach a "plateau." This plateau represents saturation of the customer with sales time. That is, his potential has been obtained and further increases of sales time will not yield significantly more sales volume.

Assuming that such a sales-response curve could be obtained for each customer, how could this information be

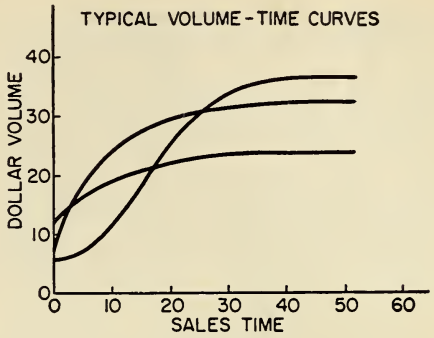


FIGURE 1

used to allocate sales time to a set of customers so as to maximize the sales obtained? This is a technical problem for which a solution was obtained. The method need not be explained here because, for reasons yet to be given, it was not subsequently used.

In order to obtain a sales-response curve for each customer, one would ideally like to collect data on the customer's annual purchase volume for a variety of different amounts of sales time spent with him annually. However, if data of this kind were available for a number of past years, it would have been necessary to make complex adjustments to account for changing business conditions. Such adjustment could introduce a good deal of error. But in this case, concern with such error would have been academic because, at best, the required data could be obtained for at most three years from only a small number of sales districts. Another line of attack was required.

To compensate for the lack of historical data, it was decided to group accounts with like characteristics and then plot for each account a point representing the amount of time spent with

it and the sales volume obtained. All of these data could then refer to a single year, in this case, 1953. Then it was planned to fit a sales-response curve to these points. If this could be done, sales time could then be allocated to classes of accounts by the method previously referred to, and then divided equally among the members of the class.

Specifically, this plan involved the following four steps:

(1) To group accounts into homogeneous classes (that is, to group accounts that could be expected to react in essentially the same way to sales time).

(2) To subclassify within each class by the amount of sales time spent with the account.

(3) To compute the average dollar volume for each subclass.

(4) To plot these averages and fit a sales-response curve to them. (See Figure 2)

Fulfillment of this plan would require two types of data in addition to that provided by the call reports prepared by salesmen: (a) information which would permit the grouping of accounts into homogeneous classes, and (b) information which would make it possible to transform data on number of sales calls into sales time.

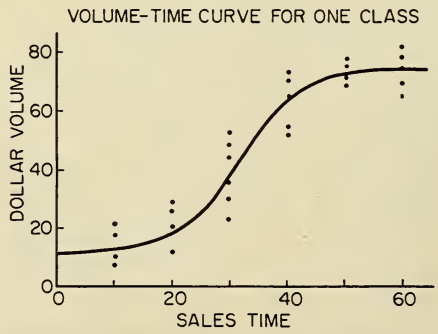


FIGURE 2

COLLECTION AND USE OF DATA

Discussion . . . yielded a list of account characteristics which one might expect to affect the way that an account responds to the time a salesman spends with it. Included among these characteristics were the size of the account, the nature of his business, his location, and so on.

Before information on these characteristics was collected on a national basis, it was decided to try out the research plan in one sales district. A particular district was chosen because it had good call reports and records and a wide variety of types of accounts. To obtain the information required, use was made of records at the Division's headquarters, the Sales District Office, and the Service District Office. In addition, the individual salesmen obtained from their accounts those data which were not otherwise available.

In order to obtain the information required to translate from calls to time spent, a survey was designed which involved daily time reports prepared by each salesman in the district for each work day during the month of June, 1953. A brief summary of the analysis of the data obtained will be given later, but it is worth mentioning here that the salesmen provided very detailed and accurate data. They were familiarized with the research plan before they were asked to do anything. Once they understood the plan they recognized

a number of ways in which they as well as the company could benefit. Their cooperation was complete.

Once the call data and information on characteristics of the accounts had been collected in accordance with the research plan, efforts were made to construct classes of accounts for which sales response curves could be found. Complex statistical methods as well as visual inspection were used in this effort, but to no avail. No method could be found for grouping the accounts so that a sales-response curve could be adequately fitted to them. Figure 3 here shows but one of the more than fifty plots of various kinds that were tried.

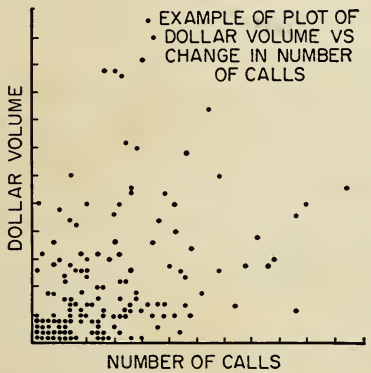


FIGURE 3

It is interesting to note that even for those accounts to which direct sale (not consignment) is made the same random scatter of points was obtained as for accounts operating on consignment.

The fact that sales response curves could not be found could have only three possible explanations, only one of which could be true:

(1) Time spent by salesmen with accounts had no effect on sales volume during 1953 and 1954.

(2) The relevant characteristics of the accounts were not taken into consideration.

(3) The amount of time being spent on most accounts during 1953 and 1954 was beyond the saturation point on the sales-response curves.

The first two of these conclusions hardly seemed justified. Consequently the research turned to a closer examination of the third. This meant, if true, that more time was being spent with most accounts than was required to get the volume of sales being obtained. If this were true, it would explain the apparently random scatter of points for accounts in the same class. That is, the points would be expected to have a random scatter about the horizontal portion on the sales response curve. The hypothesis of saturation had to be tested.

PROOF OF SATURATION OF ACCOUNTS WITH SALES TIME

Data on each account were arranged as shown in Table 1.

The numbers used in Table 1 are merely illustrative. Note that account A had an increase in both number of calls and dollar volume in 1953, as compared with 1952. Account B had an increase in calls and a decrease in volume. Account C had a decrease in calls and an increase in volume. Finally, account D had a decrease in calls and a decrease in volume.

The data for 1953 and 1954 were similarly tabulated. Graphic plots were

TABLE 1

Account	No. of Calls 1952	No. of Calls 1953	Change in No. of Calls	Dollar Volume in 1952	Dollar Volume in 1953	Change in Dollar Volume
A	20	30	+10	5,000	6,000	+1,000
B	35	40	+ 5	4,000	2,000	−2,000
C	15	10	− 5	2,500	3,000	+ 500
D	35	20	−15	5,500	4,500	−1,000
etc.						

made for various classes of accounts to show how the change in number of calls and change in sales volume for 1952–53 and 1953–54 were related. As can be seen from Figure 4 here, the scatter seems to indicate no direct relationship. It appeared from this plot that it made no difference whether more or fewer calls were made in succeeding years as far as sales volume was concerned. Statistical analysis supported this visual conclusion.

A careful study was made to find possible differences between those accounts on which more calls were made and those on which fewer calls were made. It might have been that this scatter was not random but was due to some fundamental difference in the nature of the accounts. Returning to the characteristics of the accounts that had been collected, an effort was made to find one or more specific characteristics which would distinguish between the groups of accounts on the left and right of the vertical axis of the plots. None could be found.

To check these results, the accounts of individual salesmen were plotted separately. These plots showed the same random scatter as did the composite plot. Two lists of accounts were then made from the 1952–53 change plot: those on the left of the vertical axis and those on the right. These were given to a sample of salesmen who were asked to try to find some characteristics that would distinguish between the two groups. The salesmen and others who made the effort could find no characteristic that could distinguish between the two groups.

The last procedure, however, does not prove that no such characteristic(s) exists. But we can prove this by actually observing how these accounts responded to changes in number of sales calls in 1954.

All accounts were grouped into two classes:

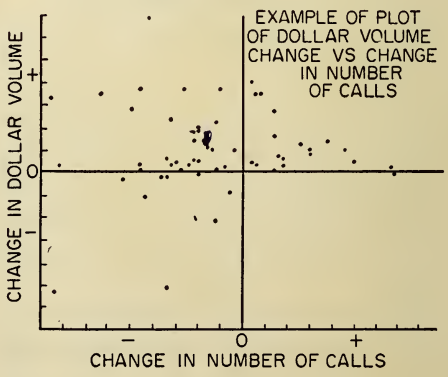


FIGURE 4

(1) Those which had an increased number of calls in 1953 over 1952.

(2) Those which had a decreased number of calls in 1953 over 1952.

It was necessary to show that the sales volume obtained in 1954 from accounts in each of these classes did not depend on whether the number of calls made in 1954 increased or decreased. Consequently these two classes were each broken into two subclasses.

1A Increased calls in 1953 and increased calls in 1954 (++)
1B Increased calls in 1953 and decreased calls in 1954 (+−)
2A Decreased calls in 1953 and increased calls in 1954 (−+)
2B Decreased calls in 1953 and decreased calls in 1954 (−−)

Then the change in dollar volume from 1953 to 1954 was obtained for each account in each class. . . . There was no significant difference in the distribution of changes in dollar volume between the four classes. This means that changes in dollar volume from 1953 to 1954 were independent of the number of calls made in 1954.

Since these results applied only to the one sales district, it was felt necessary to perform a similar study on as many other sales districts as was possible. Only two others had call reports in the required form. Data were compiled and analyzed for these districts. . . . The results were the same as those obtained in the first sales district. The composite results for the three districts are shown in Figure 5.

These figures and the associated statistical analysis show that within the range of number of calls being made,

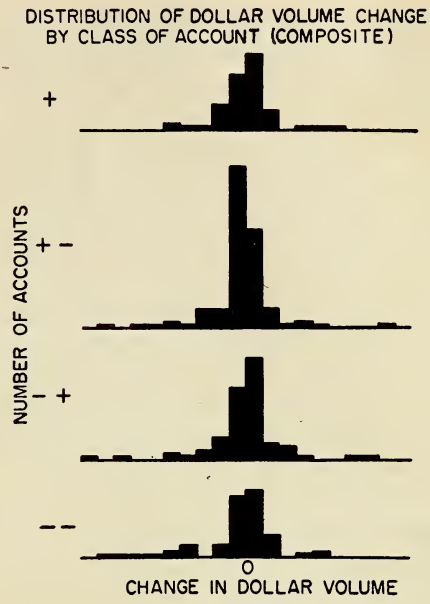


FIGURE 5

changes in sales volume are independent of increases or decreases in number of calls. This fact being established, the next question was: By how much can the average number of calls made on accounts be reduced without affecting dollar volume?

REDUCTION OF CALL REQUIREMENTS

The saturation point on the sales response curve can be estimated only if we know what the curve is. Since this curve was not known, care had to be taken not to over-estimate the possible cut-back. A very conservative procedure was adopted. It consisted of determining in which of the three years, 1952–1954, the smallest number of calls was made on each account; then a comparison was made of this smallest

number of calls with the number made in 1954. In the cases where the smallest number of calls was made in 1954, the difference was equal to zero.

This analysis showed that a considerable average reduction in number of calls could be made without affecting sales volume. This result can be put in another way: The same number of salesmen as are currently employed in the Lamp Division can carry considerable more accounts without affecting their return per account (assuming the same mix of sales and service calls, and the same mix of types of account).

This conclusion was very relevant in light of the planned expansion of the sales force. Its implications can be made more explicit as follows. On the basis of the sample of salesmen involved in this study, the increased average number of accounts that a salesman can carry was determined. Some salesmen were already carrying this many or more. A comparison of these salesmen with others carrying fewer accounts was made for one of the Sales Districts in the sample to compare for 1953 (a) their average return per call, and (b) their total sales volume. The results are

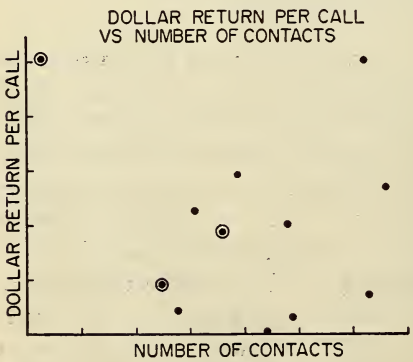


FIGURE 6

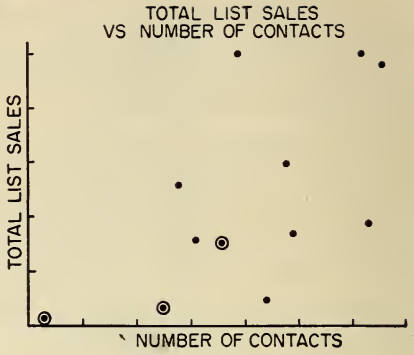


FIGURE 7

shown in Figures 6 and 7. (The circled dots refer to sales specialists who do not carry a full sales load.) These exhibits give further corroboration to the conclusion that fewer calls need be made on accounts, or equivalently, that a salesman can carry more accounts.

The planned reorganization of the Lamp Division was taken into account in determining specifically what this finding would mean to a sales district.

The analysis showed that . . . no additional salesmen were required. The operating-cost implications of this result were considerable.

LIMITATIONS ON RESULTS

The data and analysis described here showed that a reduction of calls made in 1954 had no effect on sales volume in 1954. But the question remains as to whether there is a long range or accumulative effect of a reduction in calls. Since only three years' data were available it was not possible to answer this question. It was shown, however, that the change in number of calls from 1952 to 1953 did not affect volume in 1954. This was supported by the fact

of no significant difference in change in sales volume from 1953 to 1954 between accounts with increased calls in 1953 over 1952 and accounts with decreased calls in 1953 over 1952.

This fact by itself does not justify inferences concerning longer range effects. Consequently it was extremely important to establish call records in such a form so as to detect these long-range effects if and as soon as they occur. A procedure for so doing was developed. Steps are now being taken to install this long-range control system. It is quite possible, however, that future data will indicate the possibility of further reduction in calls.

RELATED RESEARCH FINDINGS

In addition to the principal result reported above, the study generated data from which both conclusions and suggestions were derived affecting the use of sales time. These are:

(a) The breakdown of how salesmen spend their time.

(b) The allocation of sales calls to prospects.

(c) The breakdown of salesman's time into sales and service activity.

Time remains for only a very brief discussion of these.

BREAKDOWN OF SALESMAN'S TIME

Time and mileage data were accumulated for a sample of salesmen in one district. This was too small a sample covering too short a time to provide generalizations. However, the facts were sufficiently different from common assumptions as to how salesmen spend their time to justify at least one conclusion: it would be very desirable to extend the study so as to verify or refute the results obtained. For example, a comparison of travel time spent by "city" and "territorial" salesmen yielded unexpected results. The number of miles traveled per call in the country was almost double that in the city. However, traveling time per call was actually lower in the "country." Perhaps this fact is due to traffic, parking, and other time-consuming factors in metropolitan areas. If so, some ideas about the handling of suburban and rural territories out of the headquarters office should be re-examined. Or the result may be due to inefficient scheduling of calls by city salesmen. This would also suggest certain changes in assignment of accounts.

ALLOCATION OF SALES TIME TO PROSPECTS

An analysis was performed on the 1953 call data obtained from one of the sample sales districts to determine the distribution of the number of calls made on prospects (accounts large enough to justify direct solicitation by Lamp Division salesmen). A calculation was made of the number of calls after which prospects became accounts or were dropped. This suggested the question: What is the maximum number of calls that should be made on a prospect before it is dropped? An answer to this question was obtained. It showed that one would expect an 11% increase in the number of new accounts

as a result of making no more nor less than the optimum number of calls on each prospect. These results, however, were not conclusive for several reasons:

(1) Only two year's data in one district were used. Similar studies for other years and other districts should and can be made.

(2) There is a question as to whether the number of additional prospects required per year can be generated. It should be noted, however, that if these additional prospects cannot be found then it appears that the maximum should still hold and the time should be converted to calls on regular accounts. It seems likely, however, that with additional time available and an explicit policy established by management to encourage greater prospect activity, the required number of additional prospects can be found.

BREAKDOWN OF SALESMAN'S TIME INTO SALES AND SERVICE ACTIVITY

Call reports from ones sales district included a number of notes made by salesmen on the nature of their activity during their sales calls.

These remarks were summarized and listed, then classified according to whether they seemed to be primarily of a sales nature or of a service nature.

Of the calls that were described about half were devoted to sales and about half were devoted to service. It is apparent that one cannot conclude on the basis of the sketchy data available that this condition prevails throughout the country. But it suggested that further study be made to see whether or not this is true. If these conditions hold even approximately accurately over the country it suggests the possibility of relieving the salesman of some of his service functions and concentrating them in the hands of one or two persons specifically trained for service activity, thereby freeing additional and costly sales time.

In conclusion, it should be noted that the study fulfilled at least one essential characteristic of Operations Research: it raised as many or more questions than it answered. Herein lies one necessary condition for continuous progress.

✦✦✦✦✦✦✦✦✦✦✦✦✦ DETERMINING THE OPTIMUM ALLOCATION OF EXPENDITURES FOR PROMOTIONAL EFFORT WITH OPERATIONS RESEARCH METHODS

JOHN F. MAGEE

INTRODUCTION

How much should a company spend on advertising, and where should it spend it? Most companies spend money on advertising with the objective of building sales. For this reason, I would like to concentrate on the problem of getting factual data on just how much advertising and other promotional methods contribute to sales, and using this data to design an efficient sales promotion program.

The decision how much to spend on sales promotion and how to spend it is a critical one in most companies. I think it's fair to say that very few companies are anywhere near a concrete answer or basis for getting an answer to this question. Recently, for example, I sat in on a discussion with the president of one of our major companies. This company is well known as one that is sales-minded and has an aggressive marketing program. It is equally well known for the sophistication and thoroughness of its marketing research activities. For many years it has used a broad range of marketing research methods to find out who buys its products, what they think of them, what they are seeking in the product, and how many of the potential customers are reached by the company's advertising and promotional campaigns. Despite this effort, and despite the president's own long experience coming up through the marketing side of the business, he said, in effect: "I simply don't know how much to spend in advertising a major line. We spend a lot, and we've been very successful. Maybe we could spend less on this line and turn the difference to other products or to net profit; maybe we could spend more. But I'm afraid to spend less, and I don't dare spend more, so I go along with what we're spending and every few months we argue out where to spend it. Each of us has his own opinion what's the best thing to do, and we have lots of facts to back up these opinions. But I spend this money to get sales and we're missing one piece of information we need—what approach gets us sales, and how far can we go."

Daniel Seligman, in an article in FORTUNE,[1] surveyed some of the methods companies use to fix their ad-

[1] Daniel Seligman, "How Much for Advertising," FORTUNE, December 1956, p. 222.

The Frontiers of Marketing Thought and Science, *American Marketing Association, 1957, 140–156.*

vertising and sales promotion budgets. His article shows that a great deal of progress has been made since the days when advertising budgets were set with what was left over after everything else was taken care of, with no regard for the marketing job to be done. He points out, however, "Despite the elaborate procedures involved in the budget determination of large advertisers, many participants in the process retain an uneasy feeling that they are floundering in the dark. For there remains the old and tricky problem of relating advertising to sales."

From my own experience during the last decade, too many companies are in the position of the president I mentioned before. Despite the masses of marketing information that are being accumulated on who the customers are, what they think about, and what appeals to them, too many companies fix their promotion budgets on the basis that they are afraid to spend less and don't dare spend more.

I think this is an area where marketing research can and should do more. A review of recent volumes of THE JOURNAL OF MARKETING reveals a number of articles which point out the great difficulties of measuring the sales response to advertising and using these measurements. Granted these difficulties exist, although they are essentially no worse than the difficulties in other fields where progress is being made, such as in subatomic physics, weather control, or medicine. With the increasing share of product cost being devoted to marketing, I believe we have a responsibility to face these difficulties and use the same kind of ingenuity that has gone into the development of other techniques in marketing research to devise new concepts and new techniques for measuring how promotional effort affects sales in particular circumstances and for designing effective promotional programs.

The job breaks down into two parts. The first is measuring how much sales volume results from various combinations and intensities of advertising and promotional effort; the second is utilizing these measurements to build advertising programs. The first part is by far the harder.

MEASURING THE EFFECT OF PROMOTION ON SALES

Difficulties certainly exist. Some of the obvious ones are these: the influence may be cumulative over a long period of time; economic conditions may change; new products or competitors' actions may change the market; a brilliant idea may turn a poor program into a good one, or stereotyped thinking may ruin the best basic plan; combinations of approaches may make it difficult to sort out the value of the individual parts; seasonal or geographical influences may invalidate test results; the best advertising campaign in the world may be ruined by poor distribution or poor sales programs.

Harry V. Roberts, in a recent article in THE JOURNAL OF MARKETING,[2] made an excellent summary of the roles and limitations of various research

[2] Harry V. Roberts, "The Role of Research in Marketing Management," THE JOURNAL OF MARKETING, Vol. XXII, No. 1, July 1957, p. 21.

methods. He points out that rapid response to promotion and the ability to trace response makes it much easier to measure the response. Where these more or less ideal conditions don't exist, it may be felt necessary to use substitute responses, recognition or readership. He also notes that there is very little evidence to demonstrate that the substitute responses are really measuring the sales effect that the manager of a business is interested in. It seems to me that the measurement of secondary responses in marketing research that has received so much emphasis in recent years is important, but it still leaves the missing link: How many dollars in sales does the promotional effort produce? I would like to discuss a few concepts and approaches to the measurement problem that we have found increasingly useful in the last several years.

CONCEPT OF CUSTOMER STATE

The concept of customer state is in principle a simple one, but we have found it very useful in helping to clarify measurement or experimental problems as a basis for analytical work. A highly simplified example may illustrate the basic notion. Suppose a customer buys when he wants and needs a product, so that the particular time when he places an order or makes a purchase is unpredictable. However, when he does buy, if he is favorably oriented, he will buy your product; if not, he will buy someone else's. Then we might define two states—customer and non-customer—and say that a particular buyer is in one or another of these states at all times.

A buyer will change state as time goes on. A non-customer becomes a customer, and vice versa. Therefore we might set up a two-way table, such as Table 1, and attempt to measure or observe the rates of movement from customer to non-customer state, and vice versa. We might then ask a question: How does promotion affect the rate of change from one state to another? In many cases we have found that this question, the effect of promotion on rate of change from one state to another, is the most meaningful to ask.

TABLE 1

STATE TRANSITION PROBABILITIES

	Customer	*Non-customer*
Customer	.95	.05
Non-customer	.02	.98

For example, a printing company had the question how its salesmen should use their time. These salesmen called on various past and prospective customers to try and stimulate business. An analysis of a sample of past and prospective customers showed that each buyer of printing services gave about 65–70% of his total requirements to some one source and spread the rest among a variety of alternative sources. A customer did not tend to jump around from source to source with individual orders but tended to give the bulk of his business to one source until he shifted to another. As a result of these findings, the potential customers

making up the company's market were split up into groups based on estimated volume of printing requirements. The members of each group were characterized based on recent purchases as being either in the customer or non-customer state. The real job of the salesmen was seen to be to spend enough time with the accounts in each state to convert non-customers to customers and to hold existing customers.

An experiment was set up to measure the rate of movement from one customer state to another, under different levels of attention. The accounts believed to be in each state in each volume class were split into three sample sets. The salesmen were asked for a period of time to spend less than 5 hours a month with the first set, 5 to 9 hours a month with the second, and a minimum of 10 hours a month with the third. The experiment was continued and conversion rates were followed for a period of 4 months. Table 2 shows a summary of results for the conversion of non-customers to customers. The rate of conversion from the non-customer to the customer state was about 10% per month when 5 hours or more per month were spent with the account. These figures did not appear to differ significantly by volume class.

This is a highly simplified use of the state concept where only states based on potential volume, subdivided into customer and non-customer orientation, were considered. In other cases we have found the same basic concept to be applicable. For example, in banking a characterization was set up to watch the flow of small depositors between banks in a city. Here the simple breakdown between customer and non-customer was used as a first approximation, with the object of refining the state definitions later to take into account other classes of business such as special checking account or loan business. In the case of a nationally advertised household product, the same basic concept was used to analyze interview data and thereby measure the rate of change of households from non-customer to customer state, and vice versa, even though the individual household purchases were quite erratic.

The definition of customer state can get highly complex. In some cases there may be *a priori* information on customer characteristics known to have an effect on their buying habits; for example, size, location, industry, etc. Other

TABLE 2

CONVERSION TO CUSTOMERS VS. SELLING EFFORT

	Percent Converted Within		
Level of Effort	*1 Month*	*2 Months*	*3 Months*
Under 5 hours/month	0	0	8
5–9 hours/month	10	31	53
Over 9 hours/month	25	40	40

factors may have to be introduced on a purely empirical basis as a result of statistical analysis. The object is to define states in such a way that the expected purchases in a given period of time from customers in the same state will be about the same, as will the rates of transition from that state to others, and so that there will be significant differences in these characteristics among states.

We have found that the state concept is more useful where the average rate of purchasing is relatively high compared to the average rate of change from one state to another; for example, where the buyer may be expected to make a purchase at least twice during his average life in any one state. This concept obviously can be applied most directly where individual accounts can be watched. Even where this is not possible, we have found it useful as a concept to explain over-all figures which can be obtained and we expect that the measurement of state changes can have real general use as a basis for interpreting the importance of other factors studied in marketing research, such as readership or recognition.

The basic concept of state change has grown out of our work in marketing problems during the past decade. We expect that this concept will be drastically modified in the future; indeed, we hope it will be. As of now, it represents a useful concept or tool in investigating promotional effectiveness.

EXPERIMENTAL MEASUREMENTS

I am well aware of the difficulties in making field experiments to get a direct measure of the effect of promotion on customers' purchasing activity. Nevertheless, we have found experiments a very successful way of measuring these effects, if the experiments are well planned. There are a number of prerequisites for the experiment to be useful. In the first place, the experimental conditions must be carefully studied to get an estimate of the kinds of extraneous effects to be eliminated or cancelled out in the test design. Secondly, the unavoidable statistical fluctuations or "noise" must be measured as a basis for deciding how big the experiment must be to give the precision needed for a significant answer. We have found a tendency for many marketing experiments to be far too small to be meaningful. Finally, the experiment must be designed to take full advantage of the statistical design and analysis techniques which have been developed in recent decades in order to estimate joint and partial effects.

If these conditions are met, experiments can be fruitful as, first, sources of basic measurements such as customer ordering rates or rates of transition from state to state; second, tests of specific hypotheses; third, pretests of proposed operating procedures or campaigns. Just which of these three is the real purpose of the test must be clear before the test is designed, and with these precautions, many of the objections to the experimental approach can be avoided. I would like to cite one or two examples of test design which were found quite useful in particular problems.

One such test was designed to measure the influence of a manufacturer's

advertising on retail sales. This company makes a line of household appliances sold nationally through selected retail outlets. The company's line is advertised in various types of easily controlled media such as radio, television, or local newspapers. The question the company faced was: Does this advertising pay off, and what level and media concentration is the most productive? Preliminary analysis indicated that the retail outlets could be divided into 5 broad types ranging from large, full-line department stores to small appliance stores. The retail marketing areas could be grouped satisfactorily into 4 types: metropolitan, suburban, and two groups of non-metropolitan areas based on average consumer purchases. Sales statistics covering a two-year period were collected from a large sample of outlets and were analyzed to get an estimate of the size of test, duration, and size of effect needed to produce a meaningful result. From advertising costs it was possible to estimate how big an effect would have to be seen to have any significance, and the test was set up to be large enough to give a clear, positive effect in 6 months if the true effect was significant. The test was designed to analyze three alternative methods of allocating the promotional expenditures. One hundred and eight retail towns were selected as the basic sample, grouped by store and city type. The basic sample is shown in Table 3. You will notice that each group is composed of a multiple of 4 towns, of which one was assigned at random as a control, and the other three at random to the three alternative test allocations. This design made it possible to eliminate effects due to differences between towns and stores so that the sample would give a measure of the basic effect we were seeking.

During the test operation, the three test samples were subjected to promotion increased in intensity by 40% in each of three different ways. In the control group, the promotion intensity was cut by 30%. The procedure was set up to obtain sales reports month by month from each store in the sample, and these were analyzed sequentially as received. Each month an analysis of variance was made and the significance

TABLE 3

RETAIL TEST TOWNS

Outlet Class	*Area Type*				
	Metro	*Suburban*	*Non-metro A*	*Non-metro B*	*Total*
A	12	—	—	—	12
B	8	4	8	4	24
C	4	4	4	12	24
D	4	24	—	4	32
E	—	4	4	8	16
Total	28	36	16	28	108

levels obtained month by month were accumulated by means of the Chi-square test.

Within three months, the sample showed a definite positive effect, and the effect was clearly positive after six months. The test was continued several months further, to get better numerical estimates of the size of the effect under each alternative allocation, and then the treatments in the control group and the test group showing the largest effect were switched in order to demonstrate that the observed effects could be reproduced.

The test clearly took a long time to run—over a year, including time for analysis—but the proof obtained of a significant effect of advertising was the basis for an expanded program and for calculating how far and in what direction the program could go. The test design also served as a tool for setting up further tests of new media approaches as these came along.

As another example, I would like to refer to a problem discussed previously in THE JOURNAL OF MARKETING.[3] To summarize the problem briefly, the company in question makes a wide line of tools and supplies sold to a large number of accounts. The company is very well known in its field and its basic promotional plan is carried out by a group of salesmen who call on customers to explain the products and encourage their use. The salesmen had been calling on about 40% of the accounts. The accounts were rated quarterly to select those for the salesmen to concentrate on during the next quarter. Past purchasing records had been used as the basis for this selection. The original problem was to design an improved rating system to select accounts on which the salesmen should call. I will not go into detail on the analysis, since this is described to some extent in the article referred to, but I would like to describe an experiment that was made.

After some analytical work, a new selection system had been designed. This grouped accounts into some two dozen classes based on type of account and recent purchasing record, each class containing about 4% of the total number available. It was felt that the proposed system would be only a modest improvement over the existing method, since the special cases were pretty clear one way or another and it was only in the shadowy middle ground that some improvement could be made by selecting one account vs. another; but the company felt that any improvement would be highly profitable since any added sales volume would be obtained at no extra sales expense. The experimental problem was to demonstrate whether or not the new selection system in fact represented an improvement.

The normal test method the company used to try out its marketing ideas was not so different from that most companies use. They would try to select two similar areas or territories. One policy would be tried in each. At the end of a period of time, sales would be totaled in each territory, with an attempt made to adjust for differences in size, economic conditions, or other

[3] John F. Magee, "Application of Operations Research to Marketing and Related Management Problems," THE JOURNAL OF MARKETING, April 1954, Vol. XVIII, No. 4.

factors. The company had been quite skillful in working out useful tests. However, this type of comparison was satisfactory where a major difference in sales was expected but was not critical enough for the type of test needed in this particular problem.

To design the experiment, we selected a random sample of several hundred accounts. This sample was then checked against known characteristics of the company's business as a whole, to satisfy us that it was representative. After these preliminary checks had been made, each account was ranked or classified in two ways—once according to the old classification system, and once according to the new. The double classification resulted in the 4-way breakdown shown in Table 4, the vertical breakdown representing the existing selection method and the horizontal breakdown, the proposed method.

About 75% of the total accounts in the sample were classified the same way, by either system. For example, the 28% in the upper left-hand box were selected under both systems as worthy of receiving promotional sales attention. However, our real interest was in the two corner groups, where a difference in treatment would exist between the two selection methods. The proposed classification system would pick up about 1/8 of the accounts that salesmen would normally ignore under the conventional breakdown, and would drop a corresponding number from the group normally called on. If any real improvement in sales was to result from the proposed system, it had to come from these two groups. Therefore, each of these groups was split in two on a random basis. Salesmen were asked to call on half and ignore half, and this procedure was followed for a 6-month period.

The results are shown in Table 5. The gross difference between the methods was approximately 4.6%, but the test showed that a significant and substantial difference existed in the two affected groups. The test design used highlighted the particular critical question to be answered and was large enough to answer this question conclusively. Once this question had been answered, one could calculate what the

TABLE 4

BREAKDOWN OF CUSTOMER ACCOUNTS

Proposed Rating	*Conventional Rating*		
	Class A: Sales Calls	*Class B: Not Promoted*	*Total*
Class A: Sales Calls	28%	13%	41%
Class B: Not Promoted	13	46	59
Total	41%	59%	100%

TABLE 5

AVERAGE SALES IN 6-MONTH PERIOD

Proposed Rating	*Conventional Rating*	
	Class A:	*Class B:*
Class A:	$1150	$490
		$1100
Class B:	$790	$225
	$390	

over-all effect would be, and in fact the test results provided the raw material for a substantial further analysis aimed at the size of the sales budget.

These ideas and examples on the experimental approach are hardly startling, but they are aimed at raising certain points about experiments. To be fruitful, experiments must be preplanned carefully. Preplanning may be a long, hard job—indeed, it may take longer than the test itself—but this is the only way to determine the size and design needed to get a significant answer. We have found that the test size characteristically must be a good deal larger than has been typically used. This may be obtained through a larger basic sample, or in some cases the same effect can be achieved in a valid way by running the test for a longer time. By proper test design it is possible to examine multiple programs and measure the joint and partial effects. Sequential analysis methods can be used to cut short the test when a significant answer has been obtained, but if this is to be done, the tests must be designed to begin with to meet the requirements of these methods. Finally, the purpose of the test and the question to be tested must be clear and explicit in advance.

Some particular approaches or techniques in testing that we have used include black-out tests in sample areas, useful, for example, in radio, television, direct mail, or other local promotion. Randomized block designs or Latin-square test design, to eliminate unwanted effects and measure joint effects, are another extremely important set of techniques. In some cases, tests can be carefully designed on a before-and-after basis; that is, where the sample area is watched prior to the test campaign, and where the results are compared with results during and after the campaign. This is probably the most difficult type of test to control, but in some cases it is the only approach that can be used in connection with wide broadcast media.

We have found the concept of customer state particularly useful in many of these experiments. Where the concept fits customer behavior in a believable way, it gives a clearer definition of the experimental problem, and the measurements of state changes can be directly converted into sales dollars. This, for example, is one way of getting trustworthy results out of the before-and-after type of experiment mentioned above.

ANALYSIS OF PROMOTIONAL PROGRAMS

The basic concepts of analysis for designing an efficient program are straightforward. The key is having the measurement of what the effects of the program will be. With these measure-

ments in hand, the objective is to distribute the effort to get equal marginal payoffs; that is, to obtain, as nearly as possible, an equivalent return in sales for the last dollar spent on any medium or in any area. This is akin to the economist's marginal analysis. Without the measurements, the objective is simply an empty statement; with the measurements, it can be the basis for a powerful planning tool.

The techniques used for analyzing and interpreting experimental measurements range from simple arithmetic to more complex methods such as dynamic programing, discussed in the recent management science literature. The methods, however, are not the real point, because the real value of any of these methods comes from integrating measurements and marginal concepts and using these together as a basis for extension or for asking new questions.

To illustrate the use of the state concept described earlier in analysis, suppose we have in a hypothetical case set up two states—customer and non-customer. Suppose we have made the measurements illustrated in Table 6. The figures at the left give the estimated number of potential customers in each state and their estimated average value in terms of current purchases. The table at the right gives the transition rates or probabilities under normal promotion. By repeated application of the transition probabilities we can make an estimate of the size of the customer group as time goes on, and from this an estimate of business, as shown in Figure 1.

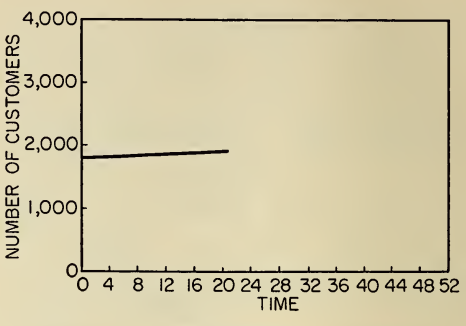


FIGURE 1

Now suppose we have a new campaign in mind and some test results produce estimates of the transition probabilities under this new campaign, as shown in Table 7. Comparison with the earlier table of transition probabilities, Table 6, shows that we estimate this campaign will cut the rate of loss from customer to non-customer state. We want to know when to start and stop this campaign.

By repeated application of the transi-

TABLE 6

ILLUSTRATIVE EXAMPLE

Customer States			*Transition Rates*	
	Number	*Value*	*To Customer*	*To Non-customer*
Customer	1890	$900	.96	.04
Non-customer	8110	$10	.01	.99

TABLE 7

ILLUSTRATIVE EXAMPLE

Proposed Advertising Campaign Transition Rates

	To Customer	*To Non-customer*
From:		
Customer	.98	.02
Non-customer	.02	.98

tion probabilities, we can trace out the effect of the campaign, as shown in Figure 2. Then, at any time when the campaign is assumed to have stopped, we would return to use of the old rates and trace out the decay that sets in, as shown in Figure 3. This is akin to setting up a system of partial differential equations and solving these for the size of the customer group in the various treatments.

By tracing out the effect of the campaign on the customer group, we can arrive at an estimate of the cost and long-term sales gain and profitability of any particular campaign. Then by trial-and-error, arithmetic, or the calculus, we can estimate how long each individual burst should last and how frequently it should be repeated. This will lead to a sawtooth type of campaign such as that described in Gerard Lambert's recent book.[4]

Where we are dealing with a problem which can satisfactorily be characterized in terms of only two states and one type of campaign, this may be a manageable approach. However, the problem becomes much more complex when we have to deal with a large number of possible customer states and many alternative campaigns. For example, the gross breakdown into customer and non-customer state might be further refined into a large number of states based on location, industry, size characteristics, or other known factors. We may have many alternative cam-

[4] Gerard B. Lambert, *All Out of Step*, Doubleday & Co., 1956.

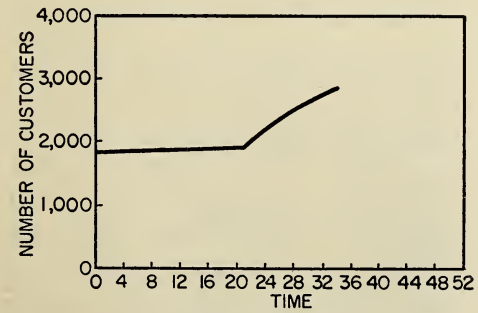


FIGURE 2

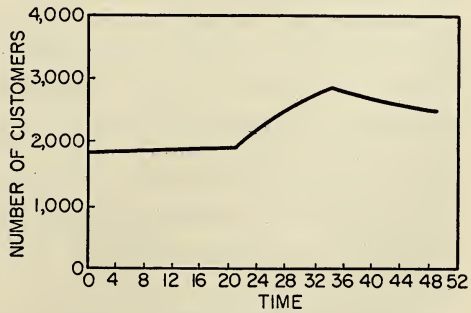


FIGURE 3

paigns to consider and these may overlap. Some campaigns may be better for specific customer states than for others. The differences may be due to media concentration, regional differences, or the influence of the campaigns on customers vs. non-customers.

Under these circumstances, balancing and selecting the most profitable campaign to run today in view of the long-term profitability of the business becomes a much more complex problem. We have found dynamic programing, originally formulated by Richard Bellman,[5] to be a very useful technique in this type of problem. For example, we have used this to design campaigns for long-term profitability where we were confronted with several dozen customer states and roughly a dozen different alternative campaigns had to be considered. By a technique such as this, the state concept and experimental results can be used together to lay out an efficient program.

As I mentioned earlier, experimental measurements are the key to this work. No matter how carefully the experiment may be run or how ideal the conditions, these are always subject to some error. Furthermore, the conditions may change, between running the experiment and running the full program. However, where a clear, quantitative description of the allocation problem has been devised (as, for example, a description based on the customer state and transition concepts), it is possible to check conclusions in advance by making a sensitivity analysis.

[5] Richard Bellman, *Dynamic Programming*, Princeton University Press, 1957.

The procedure in making an analysis of this type is to introduce arbitrary changes in the various numbers used and then observe the effect this has on the answer obtained. For example, we might re-solve the hypothetical example mentioned earlier, with the indicated effect of promotion on transition rates cut in half or doubled. Then, by working sequentially from one factor or element to another, we can search out the sensitive elements or critical combination of errors that would have a serious influence on the choice of policy. In this way, the choice of policy can be protected against undue reliance on crude numbers and the sensitivity analysis will point to aspects or elements where errors are critical and refinements in the estimates may be needed.

I mentioned an example previously where dynamic programing methods had been used to analyze a budgeting problem involving several dozen customer states and a dozen different alternative advertising campaigns. The analysis was repeated and a number of arbitrary changes made in the estimates used. For example, it was made with as much as a 1/3 cut in the estimated sales value of the individual customer states. Another change was to vary the size of the non-customer group by a factor of 10. This was equivalent to testing the policy under conditions ranging from having 5% to having 50% of the potential market. The relative weighting or discount rate applied to future values was also changed, over a range equivalent to a factor of 5. Future profits, for example, were discounted at rates ranging from

10 to 50% per year. In this way we were able to tell under what conditions the projected policy would be best and under what circumstances a careful recheck would be necessary.

CONCLUSION

The subject of measuring and allocating promotional effort is a huge one; however, progress is being made to measure the direct sales effect of promotion in various types of circumstances and to use these measurements in designing programs. I have tried to illustrate some of the approaches we have found particularly useful in our work in this area during the past decade. This work has led to some new concepts of the type indicated. I wish to emphasize that these are presented as concepts that are currently useful rather than as ultimate truth in any sense. From the interest and need which company managements exhibit, I believe these concepts, and others that will be developed by the marketing profession, will receive increasing use in seeking answers to the fascinating and vital question: How much should a company spend on advertising, and where should it spend it?

✦✦✦✦✦✦✦✦✦✦✦✦✦ SIMULATION: *tool for better distribution*

HARVEY N. SHYCON AND RICHARD B. MAFFEI

Recently a vice president of the H. J. Heinz Company, recognizing that proper warehousing was one of his company's biggest problems, asked himself these pertinent questions concerning his distribution system:

"How many warehouses should we have?"

"Where should the warehouses be located?"

"What customers should each warehouse service?"

"What volume should each warehouse handle?"

"How can we best organize our entire distribution function?"

In a firm like Heinz—with a dollar sales volume in the hundred millions, with several factories, with many mixing points where products from several factories are assembled for large shipments, with dozens of warehouses and thousands of customers—lowering the costs of distributing products to market, while still maintaining good customer service, is no easy trick. Moreover, the rising costs of distribution make maximum mileage from the distribution dollar absolutely essential.

AUTHORS' NOTE: We wish to acknowledge the work of R. K. Bennett for programing, and C. C. Beymer, I. E. Zacher, J. W. Paschke, and A. E. Buekel of the H. J. Heinz Company for their contribution to our project.

Harvard Business Review, *November–December 1960, 38:6, 65–75.*

How can answers to difficult questions like these best be obtained? The problem can be resolved through the use of simulation, one of the great advances in the science of business management developed in the past decade. Simulation provides the ability to operate some particular phase of a business on paper—or in a computer—for a period of time, and by this means to test various alternative strategies and systems. Distribution, sales and marketing, production problems—taken separately or in combination—have been solved in a remarkably accurate fashion by simulation. In the case of the H. J. Heinz Company, simulation worked with such effectiveness that a whole new approach to achieving the lowest practical costs of distribution resulted.

HOW SIMULATION WORKS

In this article, we hope to describe the way we applied simulation to solve, to a considerable extent, Heinz's distribution problems. It should be of special interest, for this simulation is perhaps the most complete, comprehensive, and accurate study of a national distribution system ever carried out.

Some readers, of course, are much more sophisticated about simulation and its uses than are others. Those who are, I hope, will bear with us as we explain each step of our method. They should remember, also, that our object here is not to write a lofty treatise on simulation theory but to share with practical businessmen, in as clear and simple terms as possible, the logic of how we went about our simulation of Heinz's distribution system. While this information can hardly be expected to enable, say, a marketing executive to tackle a simulation study completely on his own, we do hope that it will enable businessmen to understand, in general, how such a procedure could be used to help their companies test various marketing and distribution strategies.

What we are describing is a general-purpose tool—a mathematical representation of a company's distribution system. It takes into account each of the important factors involved in the operation of a distribution system: transportation rate structures, warehouse operating costs, the characteristics of customers' demand for products, buying patterns of customers, costs of labor and construction, factory locations, product mix and production capacities, and all other significant elements. These factors, taken together, make up the distribution system. Each of these elements is represented in a way which simulates its actual effect in the national distribution pattern and its effect on costs, with proper weighting and consideration given to the interrelationships among the various factors.

Since the simulation represents the essential parts of the actual distribution system, it permits the operation of the system in such a way that a whole year's transaction can be run through under close scrutiny. "Goods" flow through the system, from factory to mixing point, to warehouses, to the customer; and transportation and operating "costs" are incurred, just as they would be in real life.

But because it is only a synthetic representation, it permits the testing of various schemes for developing better distribution methods and achieving

lower operating costs. Different cost trends incurred by the alternative distribution arrangements are compared, leading ultimately to a plan of distribution at lowest cost.

For the H. J. Heinz Company, the simulation has provided a unique tool for determining the number of warehouses and mixing points which should exist in the national distribution system. It also has determined where they should be located to achieve a minimal over-all operating cost. In addition, it has provided information on how best to service the many thousands of customers by an optimal combination of service direct from factory and service from area warehouses. Further, it has given a detailed plan for allocating merchandise to given warehouses and to particular customers for each product line and from each factory. With this cohesive national distribution plan in hand, management has now proceeded to make future marketing plans with assurance of lowest actual distribution costs.

THE HEINZ PROBLEM

Heinz is typical of many manufacturers with large-scale distribution requirements. From multiple manufacturing plants across the nation, from a system of mixing points and warehouses spread across the country, the company must service all of the national marketing areas. As with many other manufacturers (both in the food and in the nonfood fields), Heinz's distribution setup has been undergoing substantial changes over the past few years.

Specific factors which have influenced traditional distribution methods are shifts in population centers and principal markets, the emergence of brand identification as a prime marketing factor, technological changes in distribution methods, the growth of large retail operations, and other changes in marketing. Added to these, of course, is the fact that the cost of physical distribution of product to market has been rapidly increasing.

As a result of these changes, the Heinz management recognized some years ago the need for a careful re-evaluation of its marketing plans and has had in process a program for streamlining and improving the marketing and distribution system nationally. More recently, it became evident that a re-examination of the transportation and warehousing system was required so that modern methods of physical distribution could be fitted to the new marketing plans in a way that would achieve a minimal over-all cost of distribution.

Heinz considered it important that a cohesive plan be developed which would combine the best features of direct plant-to-customer distribution with those of a national warehousing network. By an optimal combination of these, management hoped to minimize inclusive costs of distributing products to market and, at the same time, to maintain its policy of excellent service to customers.

GROWTH & COSTS

Problems of distribution involving both the length of distribution time and the increasing cost of getting the product to market are felt in many major segments of industry. Heinz is hardly

unique in this sense. Consumer products of all kinds—hard goods as well as soft goods, appliances, automobiles, electrical goods, clothing, the entire food industry—all are subject to increasing distribution costs.

The problem faced by the management of the Heinz company was even more complex than most. Not only were the costs of physical distribution of product to market growing, but the distribution system was increasingly being dated by a streamlined marketing program instituted at the company over the past few years to accommodate the needs of the market better and to provide improved service to customers. Included in this program were a greater recognition of the function of jobbers and distributors and a reorganization of the marketing program to build up the distributor's function in the marketing framework. Whole marketing areas had been converted from direct retailer selling to distributor areas, and, finally, an ever increasing portion of volume was moving through the distributor channels.

With changes of this nature taking place, it was inevitable that a warehouse system originally designed to handle one type of market would eventually require basic changes in order to service properly the new marketing system. Originally, the national market had been served by some 68 warehouses placed geographically to handle the many low-volume customers in the system. With the marketing structure changing, it became evident that some warehouses were located incorrectly in relation to the market now being served, and that some warehouses were simply no longer needed.

Management faced squarely the problem brought about by these changes, developed plans for the reallocation of customer volume to other warehouses, and closed some of the lower volume branches. This resulted in a reduction in the number of warehouses in the system. The company wanted to know if it had gone too far, if it had gone far enough, and, indeed, if it had retained the right locations in the system.

CHAIN REACTIONS

Management soon became convinced that the conventional methods were inadequate for analyzing (a) which warehouses to retain and (b) how best to allocate customer volume among warehouses, mixing points, and factories. In a large system of this kind changes in distribution pattern—even at the local level—tend to have chain reactions throughout the national system. A change which may appear to yield a lower cost of operation at the local level can, in fact, cause an increased cost of operation when all relevant costs are considered on a national basis. For example:

> When a warehouse is placed close to a given customer, the cost of delivering merchandise to that customer may well become lower. But the over-all effect on cost of transportation of merchandise from the various factories to the warehouse, and costs of delivery to other customers still farther way—all these, combined with the cost of operating the given warehouse, make the problem complex indeed.

What was needed, management decided, was a bold new approach to studying the distribution system as a whole, on a national basis. The many

interrelated costs, the many source points and many thousands of customers throughout the country, all had to be taken into account in establishing a distribution pattern of warehouses and mixing points which would yield the lowest over-all cost.

SIMULATION REQUIREMENTS

Our complete representation of Heinz's complex, high-volume national distribution system had to be detailed enough to handle each of the thousands of customers in the Heinz system. Specifically:

It had to take account of each customer's order sizes, his ordering patterns, the various types of shipments he receives, and his product mix.

Provision had to be made for handling the costs of the various kinds of shipment made—i.e., carload, less-than-carload, truckload, less-than-truckload, and various shipment sizes within the lower classifications.

Variation in warehouse operating costs —i.e., labor costs, rentals, taxes for different geographic areas—had to be considered.

The many different classifications of products which Heinz manufactures, the alternative factory source points for each of these products, and the factory capacity limitations on each—all had to be examined.

Finally, when such a representation was designed, it had to be in such form that it could be synthetically operated, using real operating figures, for a year's time, over and over again.

In this way, various configurations of warehouses and mixing points could be tried so that costs might be observed for different conditions, and the lowest cost pattern achieved. And since the number of transactions required for one year's operation of the national system would be so great, the representation had to be in such form as to be operable on a high-speed computer.

LOGIC OF SIMULATION

A distribution system exists in order to link production activity (which, of course, can*not* exist everywhere) and consumption activity (which *does* exist almost everywhere). A company interested in studying its warehouse location problem could start by specifying where production takes place and where the majority of its customers are located. It could, initially, assume arbitrary locations of warehouses. If proper cost information, consumption information, and production information are available, then the costs of distribution associated with a given *assumed* configuration of warehouses could be determined. These results could be compared with costs accruing under other assumed configurations.

This idea is simple enough, but the question that immediately suggests itself is this: *How can sufficient detail be designed into such a representation to provide genuine assurance that the lowest-cost distribution plan developed on paper would be realized during actual operations?*

This question is not only legitimate but is of crucial importance when analysts talk of studying and simulating systems. The answer lies in the nature of the simulation developed. Properly designed, the simulation takes account of all relevant aspects of the problem as they interrelate with one another,

and operates much as the real system does.

It might be said that simulation, in providing the means for testing the various alternative courses of action available, simply evaluates all of the "What if?" questions frequently asked. It tests those things that businessmen would like to try if time, money, and manpower permitted. For example:

> Without simulation, Heinz could have done a cost analysis for each of a number of distribution systems under various assumptions as to sales patterns. Each such analysis would have been rather costly to conduct. Analysis of a single national distribution configuration, yielding one year's operating results, required some 75 million calculations by the computer—and these were performed in less than one hour. By conventional methods, this would have taken two clerks almost 50 years!
>
> Further, the number of alternatives which could be economically examined in this way would have nowhere nearly exhausted all of the possibilities. Analysis of just 20 such possibilities would have required 2,000 clerks working one year! Moreover, management could not feel any great confidence that its final decision was "correct," because of the probability of human error and the great passage of time during which things would change.

BASIC FACTORS

To assure that the results of this study would be meaningful, we first had to specify the characteristics of Heinz customers and factories. Each customer's characteristics were specified according to:

Geographic location.

Order sizes and frequency.

Volume of purchases.

Variety requirements.

And each factory's characteristics were specified according to:

Geographic location.

Production capacities by product line.

Product mix.

Between these two basic factors—customer location and needs, and factory location and production characteristics—lies the distribution system. The problem, then, becomes one of determining the number, size, and location of warehouses and additional mixing points which would properly serve customers at a minimum cost nationally.

In a dynamic distribution system of this type many forces exist which influence warehouse number, location, and size. The nature of each customer order—its product mix, its timing, the effect of special promotions and pricing policies on customer ordering and stocking, and other factors—all are influential. In similar fashion, every applicable freight rate from each geographic point to every other geographic point, the freight rate "breaks," and similar transportation specifications have their effect. The cost of operating a warehouse at each potentially alternate location has its influence. Finally, the precise product mix of each factory, along with the capacity limitations by product line, affects warehouse location and the cost of distribution.

To evaluate properly each of these characteristics—for the many thousands of customer orders, for the thousands

of alternative sources and routings possible, for the multiplicity of alternate possibilities of warehouse and mixing point configurations—it was necessary to construct a mathematical "model" of the distribution system. Adding high-speed computing ability completed the requirements necessary to a solution of this problem.

Included in this "model" are all the essential parts of the distribution system which influence warehouse location. But which parts of the distribution system are essential and which are not?

The answers were found only after considerable research into the actual distribution records of the company. Specifically, these are the factors that had to be taken into account in setting up the model:

1. How frequently customers order, how much they order, what they order, where they are located, and how they prefer to take receipt of ordered goods.
2. The kinds of goods that can be supplied from any given factory point, the quantities that can be supplied, and the location of the factories.
3. The relationship between shipping rates and points of origin and destination, for truck and rail transportation, and for different types and sizes of orders.
4. The relationship between total handling costs and total volume handled at warehouses and mixing points.
5. The knowledge of where these relationships differ, so that adjustments to cost and volume estimates might be made.

Once this information was obtained, we then had to establish some basic working definitions of the terms *customer, factory, warehouse,* and *carrier.* And for our work at Heinz, the definitions had to be in precise numerical terms.

A. What is a customer? In terms of distribution requirements, a customer can be defined according to the following criteria:

By specific geographic location.

By business type (that is, whether it is a chain, distributor, wholesaler, jobber, vendor, or a hotel and restaurant distributor).

By product-mix consumption pattern—As a result of a thorough search of internal product records by customer account, some 50 different consumption patterns were isolated. Each customer in the national system was assigned that pattern which best reflected his product usage.

By frequency, quantity, and patterns of ordering—Each customer has his own way of ordering and of taking inventory. Some may accumulate requirements and take only large quantities; others may order frequently and in relatively small quantity shipments. The option is theirs. It was considered essential to reflect each customer's ordering patterns explicitly since this aspect was felt to have a great bearing on the distribution system.

By proximity to Heinz's various warehouses.

B. What is a factory? In terms of distribution requirements, a factory can be defined like this:

A geographic location that produces various company products.

A product-mix pattern—Not all factories turn out every product that the company makes. Therefore, in the Heinz study we defined a product as having both food and location characteristics. . . .

Product #1 was defined as a class of

varieties that were produced only at Factory #1. Product #2 was a member of a class of products that were produced at Factories #1, #2, #3, and so on for Products #3, #4, and #5. This process of classification covered all the products in the Heinz line.

A production capacity pattern—Capacity by product lines is difficult to conceive and to measure in a multiproduct, multiequipment plant. Therefore, in order to get an idea of system costs, production restraints must be imposed. This was done in the Heinz case, and a production capacity pattern was established for each of the factories in the system.

A controllable source—It is important to note that management has within its direct control the power to expand or contract capacity, to add or subtract product lines, and so on.

A cost area for transportation purposes—That is, the costs of shipping to an area of 100 miles around Chicago will probably not be the same as the costs of shipping to a 100-mile area around Denver.

With the simulation now completed, management has the means for testing various changes in production or marketing strategy. It can better answer questions as to whether additional factories should be allowed to produce a given product, or whether economies would result if certain products were removed from a given factory's production schedule. Let us take a simple example.

In the case of Product #1, which currently is produced only at Factory #1, we might wish to know what over-all costs would be if we were also to permit Factories #2 and #5 to manufacture Product #1. How do we find out? Simply by adding production capacity for Product #1 to our simulated Factories #2 and #5, and once again operating the system within the computer. A new cost of distribution will then indicate whether such a change is desirable.

We find this an excellent way to bring to management's attention the potential savings to be had under various assumed conditions.

C. What is a warehouse? For the purposes of our study, we defined a warehouse as:

A geographic, gathering, sorting, and redistributing point—A warehouse performs work, owns or rents space, employs people, pays taxes, and in general accumulates costs. In a study of this kind, it must be assumed that any geographic area in the United States is a potential location of a warehouse. And costs differ by geographic area.

A cost accumulation point—Geographic area cost differentials must be recognized in any study of warehouse numbers and locations. In a simulation of a national system it is most feasible to build into the model cost-adjustment factors by geographic area for the various cost elements.

For Heinz, the country was divided into a large number of "cost areas," and cost-adjustment factors were developed by area for warehouse labor, taxes, rentals, or depreciation. Hence, when a warehouse was placed in a given geographic area, its cost of operation was computed using the local area costs. When, in the study, the same warehouse was moved into another geographic area, the cost of operation was computed using the new area's costs. By this method a given warehouse might be more or less attractive for serving certain customers, based not only on the transportation cost for serving them, but also

on the operating cost of that warehouse versus other warehouses in other areas which might have different operating cost structures.

D. What is a carrier? For our purposes, carriers were defined as:

Either a trucking firm or a railroad.

A cost for moving goods between geographic points; in effect, a geographic movement-cost relationship—It is extremely difficult to analyze transportation rate structures. Yet when we wish to determine distribution costs, we must draw a pattern of freight costs which accurately reflect the national rate structure with all its differences depending on size of shipment, type of carrier, and other important factors.

Nevertheless, after much effort, basic regularities governing rate structure have been determined and have been made part of the simulation now accomplished. When, after careful consideration in the study that we made of the Heinz distribution system, it was decided to use relationships rather than point-to-point costs, we did this with the assurance that there was genuine regularity, and that the results were indeed authentic.

It is worthy of note that the transportation rate structures are frequently further complicated by other factors. Some customers cannot or will not accept certain types of shipments; some cannot accept rail. Some, for reasons of their own inventory policies, will not take shipments above certain sizes; others prefer not to accept small shipments. All these factors complicate the analysis problem and make necessary the use within the simulation program of fairly complex rate structure relationships based on type of shipment, shipment sizes, geographic area, and other pertinent factors.

PROGRAM CHARACTERISTICS

A sequential flow of subcomponents and components representing the flow of raw materials and finished goods through the many processing and transfer points forms the simulated distribution system. Basically, customers place requirements on a system and the system responds. Demand thus usually "explodes" backwards through the production and procurement system. But this backward explosion of demand will vary somewhat in the channels used among different industries. For example:

In the distribution of automobiles, customers in an area place orders first with dealers. The dealers then refer orders back to an assembly plant which, in turn, places demands for subcomponents back on suppliers and factories.

In the case of food and pharmaceuticals, on the other hand, customers place packaging demands back on warehouses or manufacturing plants of container companies. These companies then place orders with suppliers of raw materials or with other manufacturers.

In the Heinz system, as we conceived it, customers place orders with the company and the company responds by delivering in one of three basic ways, depending on which way or combination of ways offers the least cost. These three are:

Direct shipments from a given producing factory to large customers.

Shipments from various factories to a so-called mixing point located at a factory and then to customers.

Shipments from factory points to a warehouse and then to the customer.

WAREHOUSE FUNCTION

Let us focus our attention now on the warehousing aspect of the system and ask a basic question: Why do warehouses exist? Under what conditions might they be unnecessary? If all customers were very large, if all of them gave sufficient lead time when ordering, and if all factories produced the full line of the company's products, then all shipments to ultimate consumers could be made directly from the factory. Thus the main reasons why warehouses exist are that customers are not large enough to warrant direct shipment, and do not all give sufficient lead time when ordering, and that individual factories are not always full-line.

Since our objective, however, is not only to determine the number and location of warehouses but, even more important, to design a total distribution system which will operate at lowest total cost, it is necessary that we assign customer shipment volume to its highest distribution classification. Thus:

> If a given customer's volume is such that he qualifies for shipments direct from producing factories, and if he is willing to accept shipments by this method, then we should ship that way.
>
> Similarly, if part of a customer's volume might most economically be shipped from a mixing point, then this method is proper.

Hence, only after other volume has been allocated do we consider warehouses for shipment. And our simulation must be designed to make these determinations automatically.

DIRECT SHIPMENTS REMOVED

Some customers in the system can take part of their total demand in direct shipments. Because shipments direct from producing factories to customers bypass the mixing point and warehouse system completely, direct shipments of this kind, as Exhibit 1 shows, have no effect on the optimal placement of mixing points and warehouses. Therefore, all direct shipment customers are eliminated from consideration when we are concerned with warehouse location.

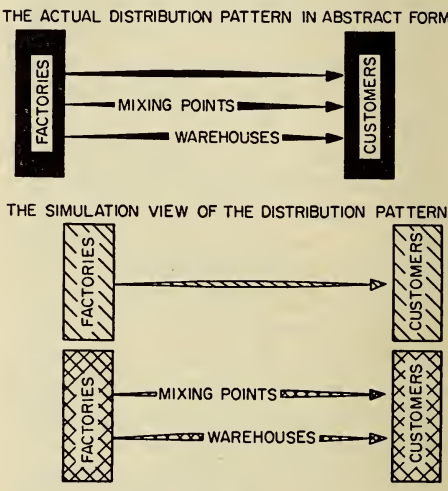


EXHIBIT I

ACTUAL AND SIMULATED VIEWS OF DISTRIBUTION

Similarly, direct shipment volume is removed from the order patterns of those customers who take only part of their demand in direct shipments. That is, when the computer found a customer whose volume of given items was large enough to take shipments direct from producing factories, it made

separate record of the volume so delivered and listed only the remaining volume for delivery from warehouses or mixing points.

This adjustment needs to be done only once and can be done by simulation. After removal of direct shipment volume for every customer, a single run on the computer will make available the resultant consumption patterns of the national system. This, then, is the information used to study the warehouse location problem.

THE COMPUTER PROGRAM

When talking in terms of large-scale computers, we should bear in mind one thing. Although computers are frequently called "electronic brains," they are by no means thinking machines in the human sense. A computer is merely a mathematical "beast of burden" which will do only what it is specifically told to do. But it does its assigned job with a speed and accuracy far beyond any other known means, human or mechanical. Instructions to the computer, therefore, must be precise and in detail. These instructions on how to proceed are called the computer "program."

In concept, the program for the simulation described is quite simple. Stored on tape is all information relating to transportation, handling, and delivery costs, geographic adjustment factors, factory locations, factory production specifications, and the volume remaining after elimination of direct shipment volume. Even the program itself is stored on tape.

The basic process is to vary warehouse configurations and to observe and compare the resultant effects on distribution costs. To do this, we must compute in detail the annual costs for operating the proposed nondirect distribution system for a year. Included are such costs as those for each of the warehouses and mixing points, for all shipments (both from factories to warehouses and warehouses to customers), and for each of the several thousand customers. Further, these costs must be broken down for each product class and each type of shipment.

OPTIMIZING THE SYSTEM

Now the simulation is ready to accomplish its twofold objective: (1) to enable management to close in rapidly on the number and approximate locations of warehouses which will achieve lowest costs of distribution, and (2) to discover where changes can be made in warehouse locations which will lower costs still further.

When the simulated one-year operation of a complete warehouse configuration has been completed and all costs computed, the following results are shown in detail as computer output.

Costs are shown for all pertinent items indicated in accounting terminology familiar to management. For each factory, warehouse, and mixing point, there are three major categories of distribution costs determined in the simulation:

Costs of direct shipments, factory to mixing point, mixing point to customer, and factory to warehouse shipments.

Costs of operating both mixing points and warehouses at specified locations.

Costs of shipping from warehouses to customers.

All these costs are further classified by size of shipment, and include a volume-by-product-line breakdown for each warehouse.

Customer-warehouse affiliations are given so that accurate service areas are built up for each warehouse, mixing point, and factory. All this is based on a lowest cost for operating the entire distribution system.

In short, a great deal of useful information about any distribution configuration is provided by the simulation developed.

Exhibit II may prove helpful at this

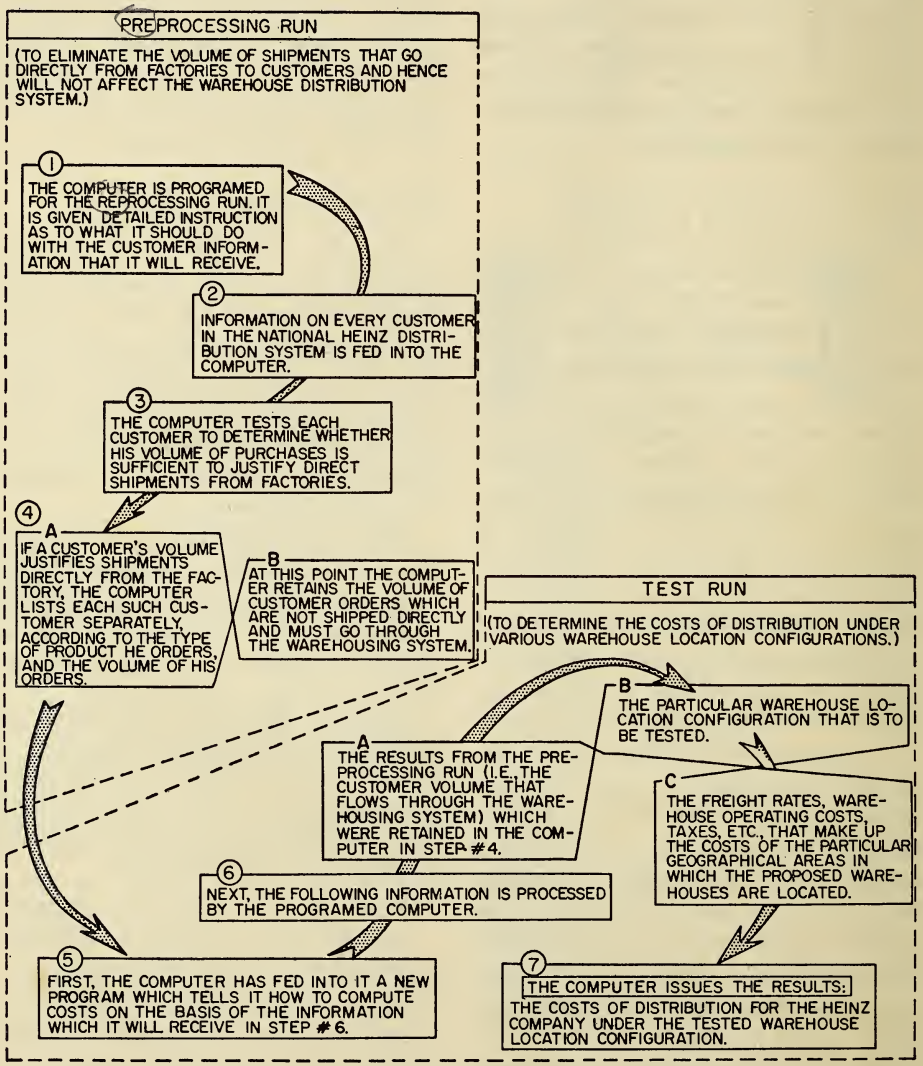


EXHIBIT II

HOW THE SIMULATION TESTS A PARTICULAR WAREHOUSE CONFIGURATION

point as a summary of the step-by-step action we took in using the simulation to test one particular warehouse pattern. A similar process, you will realize, took place for each warehouse and mixing point configuration we tested.

With the design of the simulation described, and the substantial research performed, the results showed a very distinct cost minimum. The cost of distribution which was minimized was that broad concept which includes costs of transportation between Heinz factories and warehouses, costs of operating the warehouses in various locations, and the cost of final delivery to the Heinz customer.

The results showed clearly that for the distribution requirement of the H. J. Heinz Company (a given optimal configuration of mixing points and warehouses, with given locations, and serving given customers in accordance with prescribed procedures) a lowest over-all cost of national distribution would be realized. The results are logical and attainable.

An area map is shown in Exhibit III to illustrate hypothetical warehouse locations obtained. For Heinz, a complete national map of actual warehouse locations recommended was drawn to provide a visual identification of the new distribution system. In addition, we were able to draw precise warehouse-to-customer assignments, specifying which warehouse would best serve each customer in the national system for each type of shipment received. Further, where shipments were large enough, we specified which mixing point should be used and, if a direct-from-factory customer, which factories should ship. These precise customer

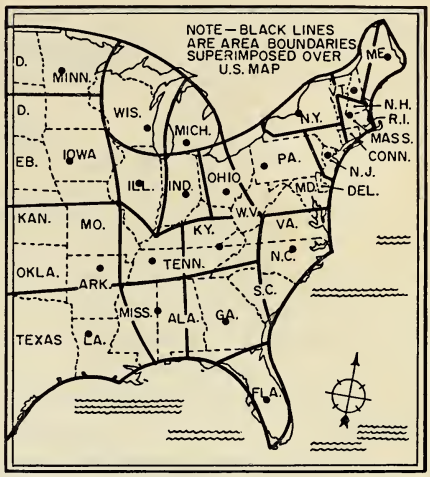


EXHIBIT III

HYPOTHETICAL EXAMPLE OF CUSTOMER ASSIGNMENTS AND WAREHOUSE LOCATIONS OBTAINED FROM THE SIMULATION

assignments then indicated exact area outlines for each warehouse. Samples of warehouse area assignments are shown in Exhibit III.

EVALUATING THE RESULTS

While the simulation does indicate such things as the optimal configuration of warehouses in exact figures, it is by no means a substitute for the judgment of management. As shown in Exhibit IV, for example:

> The results indicated clearly that for most efficient operation it was necessary to have M warehouses, but that it did not matter much, from a dollars-and-cents point of view, whether there were as few as L or as many as M. Costs were about equal under either alternative or any alternative in between. From the point of view of customer service, we recommended M, although a good case could have been made for some number in be-

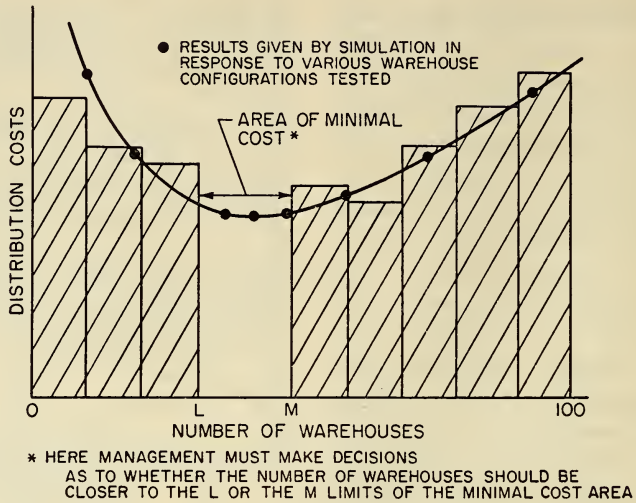


EXHIBIT IV

ARRIVING AT THE NUMBER OF HEINZ WAREHOUSES THAT WOULD MINIMIZE TOTAL DISTRIBUTION COSTS

tween. L and M differed by some few warehouses.

In making choices within ranges such as this, solid judgment, experience, and knowledge of local conditions in given areas come to the assistance of the simulation. And, as we did in the Heinz simulation, trends in the industry and economic arguments beyond the scope of the simulation should be examined carefully to determine whether the recommendations of the simulation were indeed proper ones.

OTHER USES

The method for performing the Heinz study has provided great facility for studying other aspects of the business which are at least as important as the development of an optimal distribution system. When the simulation was developed, it was thought important to build a general-purpose tool, one which management could use at any time, future as well as present, to study questions of major concern. It was not, however, until the simulation was designed that it was fully realized that the tool provided such facility for studying a wide range of perplexing management problems. Specifically:

Distribution cost studies—Customers can be separated by areas, types, shipment sizes, salesmen, type of carrier, channels of distribution. We could get estimates of distribution costs on the basis of each or any combination thereof.

Locational studies—The number and location of factories could be changed, for example, rather than altering the warehouse configuration. Then, too, the effect on the company's operations of a sudden shift in customer type or location could be studied.

Studies related to products—The product mix at each factory can be changed arbitrarily to observe whether adding product capacity would change distribution costs appreciably. Similarly, customer consumption patterns can be altered to see what effect such changes will have on distribution costs.

Studies related to time—Customer data can be altered in order to reflect gross annual volume changes by product line. These data would then be used to determine distribution costs. Thus, it can be seen what effect proposed changes in sales policy, prices, or new products would have on customer purchasing frequency, order size, or volume. The possible effect on distribution costs and on profitability can be estimated experimentally.

LIMITATIONS IN USE

While the simulation is a remarkable management tool, it does have its limitations:

(1) Resources can only be stretched so far. Some compromises obviously have to be made, although any compromise that might seriously reduce the meaningfulness of major results must be avoided or the project may be worthless.

(2) The technical characteristics of the equipment set bounds. The program we used was written to be fast and versatile. This meant that much had to be stored in the computer's internal memory, and in a problem of this size it does not take too long to jam up against a 32,000-word ceiling.

(3) The accuracy and adequacy of input information impose limits on the program. If any one maxim developed out of this study, it was this: Know your customers (i.e., get control over your input data). Results are only as good as the data that are used to create them.

CONCLUSION

Great advances in the science of business management have taken place in the past decade on a scale unprecedented in the history of business planning. Perhaps one of the most useful techniques developed is simulation. Simulation provides the ability to operate some particular phase of the business, on paper or in a computer, for a period of time and by these means to test various alternative strategies. Distribution, sales and marketing, production, or even all in combination, can yield to this new science.

Every major decision-making executive has long wished he might, by some means, test the various alternatives open to him before making a final decision on a complex problem. With the development of simulation, a major breakthrough has been achieved in providing this insight into the future. To be able to test many alternative courses of action and to obtain documented evidence of the operating results of such proposed action, places in the hands of the aggressive businessman a tool of inestimable value.

The use of high-speed computers is a principal element contributing to the feasibility of examining the various alternatives. In order to put the problem on a computer, it is first necessary to express the problem and the characteristics of distribution in mathematical form, that is, to construct a mathematical model of the distribution require-

ments. Once it has been designed, however, the model may be looked on as a form of capital investment which makes possible economies in analysis, both present and future.

The importance of this accomplishment must be considered not only in the light of more profitable distribution, but also as a basic adjunct to policy determination and the study of profit achievement. Regardless of whether a company is in hard goods or soft, in food manufacturing or electrical appliances, in consumer or in industrial products, it can conduct, through simulation, a wide range of basic studies. Area profitability, product line and type of customer profitability, the effects of pricing "breaks," all these can be studied just as the factory and warehouse location problems have been by the Heinz company.

Certainly, though, we do not mean to imply that all the major logical difficulties of simulating a distribution system have now been solved. The remaining tasks of embellishment and increased accuracy are great. But we do want to assert that such simulations are powerful tools. In the business of the future, perhaps, every company will have one.

C

Operations Research in Financial Management

PROBABILITY STATISTICS IN ACCOUNTING

A. C. ROSANDER

THE CASE OF THE STOLEN STOCKINGS

Probability statistics is a strange science. It is strange because it has a checkered past, its development having been associated with gamblers on the one hand and with college professors on the other. It is strange because, due to its association with gambling casinos and ivory towers, it has lain dormant as an applied science for over two centuries. It is strange because even though it is a powerful and versatile science with revolutionary implications for industry and government, for research and management, these implications have been recognized only during the past 30 years. Finally it is strange because it removes three common activities—sampling, estimating, and inferring—from the area of common sense and intuition and puts them on a scientific basis.

Despite these recent developments applied probability statistics has not yet received general recognition; indeed its role is not yet understood by any large number of specialists or administrators. As an illustration of the difficulties encountered in trying to apply probability statistics to the prob-

Industrial Quality Control, *May 1955, 26–31.*

lems of management, let us consider a story which appeared in *Nation's Business* for April 1952, a story which a writer of detective fiction might call "The case of the stolen stockings."

It seems that the owner of a large hosiery factory had reason to believe that 100,000 dozen pairs of nylon hose, worth about a million dollars, were vanishing from his plant every year. He hired detectives, put automatic recording devices on the machines, had foremen quizzed, but could discover nothing. Finally he called in a firm of consulting psychologists, apparently believing that they, as specialists in human relations, could find what the others had missed.

The psychologists came. They set up motion picture cameras and tape recorders. They studied films and records. Still nothing significant could be discovered. Finally one of the psychologists began to ask questions about how certain figures were obtained. This questioning marked the turning point in the case because it eventually revealed the origin of the inference that the hose were being stolen.

Briefly the facts in this case were the following: Management based its anticipated annual output on a test run in which one of the better machine operators made a number of sample stockings. Apparently this operator used about the same amount of material, 207 grains, for each stocking, so that this figure was used to estimate the number of pairs of hose to be expected from the stock of yarn purchased during the year. It was thought also that since semi-automatic machines were being used, there would be no variation in this value of 207. Rough tests made by the psychologist in the factory showed however that there was not only variation in this value of 207, but that the best operators tended to use less yarn than did the poorest operators. Hence the value of 207 was much too low. Actually the psychologists showed later that there were about 50 ways in which the operator could affect the quality of the product from this semi-automatic machine.

These facts showed what had happened. The estimate of yarn per pair of hose was biased; it was much too low. This meant that the estimated amount of annual production from a given stock of yarn was much too high. The answer was clear: a million dollars' worth of hose had not been stolen; they had never been made! The million dollars was the added cost that the owner had to accept because the average performance of all of his machine operators was not as high as that of the girl who had made the sample stockings.

This example illustrates several significant but common situations. The solution to this problem was neither legal nor psychological; it was statistical. The nature of the difficulty was discovered more or less accidentally. It is very difficult for management and non-statistical personnel to recognize a statistical problem, or the statistical aspects of a problem. Neither the foreman nor the higher level administrators understood the concept of variability as it applied to estimates, measurements, machine operations, and human performance, nor the relation of sampling to this variability. Experience, intuition, and common sense offered no

solution to this problem; what was required was a knowledge of the principles of sampling, estimation, experimentation, and inference. Such a solution would have been a routine matter for the experienced applied probability statistician. He would have recommended at least three changes: the use of a designed experiment to obtain unbiased estimates of material consumption per unit of product obtained; the use of statistical quality control charts for the purpose of controlling the consumption per unit of product for each of the machine operators; and the use of the correlation between yarn input and the hose output as an additional method of control—in other words, the use of input-output analysis.

CAN PROBABILITY STATISTICS BE APPLIED TO ACCOUNTING?

In the example just cited one may wonder why the owner did not call in an accountant since money losses appeared to be involved; one possible reason may be found in his strong feeling that the cause was theft while the other may be that he was as unfamiliar with cost accounting as he was with probability statistics.

It is easy to see, however, that if the accountant had based his estimates of costs on the production department's sample runs he would have been a victim of the biased sample, the same as the others were. This brings us to the question as to just what is the relation of probability statistics to accounting. As a point of departure we shall cite and discuss four statements made by accountants and auditors relative to the role of probability statistics in accounting. This seems appropriate in view of the misunderstanding which exists relative to the applicability of probability statistics.

In discussing the application of probability sampling to the test check of the independent auditor, Kohler makes the following statement: [1]

> Examinations of accounts are, for the most part, dependent for their effectiveness on sampling or testing. From the meaning and accuracy of a number of entries in an account, the meaning and accuracy of the whole account are judged. When errors and other irregularities are discovered, a more thorough examination is required. Sampling skills are not easy to acquire. They cannot, for the most part, be based on the mathematics of statistics; the "universe" (or population) from which the accountant samples is too intricate and unstable, and it is generally impossible to determine relevant possibilities in advance.

The first two sentences in the foregoing quotation are a good statement of the role of sampling in auditing financial accounts, providing by "sampling" that we mean "probability sampling." For if we use probability sampling we know precisely the risk which we are taking, due to sampling, when we infer the meaning and accuracy of the whole account from the data which we obtain from the sample. If we use "judgment" sampling we do not know what this risk is, nor how to make es-

[1] E. L. Kohler, *Auditing—An Introduction.* Prentice-Hall, New York, 1947, p. 9. Kohler has edited "A Dictionary for Accountants" (Prentice-Hall, 1952), a dictionary which contains many basic concepts and methods of modern probability statistics. The quotation should be interpreted with this in mind.

timates from the sample to the population.[2]

The statement that sampling cannot be based upon probability because the population or universe is too intricate or unstable is not borne out by the facts. The author indicates that the sample is to be used to judge the meaning and accuracy of the "whole account." Actually the "whole account" is the universe or population as the context clearly shows.

The belief that their subject matter is too intricate and unstable to be amenable to probability statistics is not uncommon among specialists, experts, and administrators. The economist thinks the economy is too complicated, the biologist thinks the human organism is too unstable, the sociologist thinks human organization is too complex to warrant any valid applications of probability statistics. The business man often thinks the same way; "statistical quality control may be applicable to someone else's problems but it is not applicable to mine; my problems are different." The chemical manufacturer says that statistical quality control may be applicable to mass production industries but it is not applicable to batch production nor to continuous process manufacturing.

[2] A judgment sample is one in which the sample units are selected on the basis of personal judgment or choice rather than by a random process; hence the principles of sampling, estimation, and inference based upon random or probability selection are not applicable to it. Judgment samples are usually selected on the basis of what units or elements one thinks are typical or representative; e.g., accounts transacted during a typical month.

Yet for over two decades we have been applying probability statistics to the problems of economics, biology and sociology. We have been applying it to problems of business and manufacturing and to chemical processes. And in a few isolated spots we have been applying it to the problems of accounting, taxation, and finance.

In this same connection consider another statement made by a certified public accountant.[3]

> Until vastly more information is available and until it is carefully analyzed from statistical as well as auditing points of view, public accountants can hardly expect to be able to put statisticians to work designing samples. Prior to that step there must be a codification and analysis of present criteria in the selection of audit samples and the drawing of inferences from them. . . .
>
> Serious and widespread research in this important matter should be undertaken by the accounting profession. . . . Such research would consist, in the early stages, principally of organization and analysis of the facts disclosed by working papers of past and current engagements. Only after such exploratory work is carried out can there be serious and fruitful consideration of the general application of statistically designed samples to audit test-check problems.

This position does not take into consideration a number of important points which are characteristic of applied probability sampling. Efficient sample design requires knowledge of certain characteristics and risks, such as means,

[3] Robert W. Johnson, "Statistical Techniques for Auditing Need Deeper Study," *The Journal of Accountancy*, September 1953, pp. 336–340.

variances, frequencies of occurrence, and the direction and magnitude of errors. This knowledge can be obtained best from probability samples rather than from working papers and other data which are derived from judgment samples subject to unknown sources of bias. It is quite likely that much of the information and analysis mentioned in the quotation, will be of little or no value in designing probability samples.

Fortunately probability sampling can begin without the accumulation or analysis of large masses of information; the necessary knowledge is accumulated as we go along and this knowledge can be used as the basis for an improved sample design. Deeper study is needed in applying probability statistics to accounting, but study and research and analysis are not going to be fruitful until accountants and applied probability statisticians get together, discuss these problems, and actually test probability methods in auditing practice.

It should be emphasized that the characteristics of financial transactions and records as met in accounting offer no serious logical or technical obstacles to the application of probability statistics. The fact that a characteristic may exist in a heterogeneous state, that it may have a highly skewed frequency distribution, that it may occur rarely or sporadically, that it may vary with time, that it may be associated with a process, that it may be dominated by human rather than by machine activity, that it may occur in varying proportions, that it may appear to be unstable or intangible—is nothing new to the experienced probability sampler. Indeed, the technical problems encountered in the field of accounting are similar to those found in other and widely divergent fields.

The accountant has to make an inference about a finite number of business transactions conducted during a fixed period of time such as one year. The population or universe which he samples may take one of many different forms. It may be books, or pages within books, or items on a page. It may be pieces of paper such as invoices, checks, vouchers, bills. It may be record cards such as are contained in an index file or a punch card file. In a large business it may even be departments, stores, or warehouses. It may be physical assets in one form or another. In a company engaged in mining or selling precious ores, the accountant may be very much interested in how these ores are sampled and tested. In all of these situations a population or universe does exist, it can be defined, it can be circumscribed, and it can be sampled according to the principles of probability statistics.

Let us consider another aspect of the problem which is suggested by the following statement issued by the American Institute of Accountants: [4]

> The extent of testing in any audit is decided by the CPA in the light of his best independent judgment as to the amount required to constitute a fair sampling of the record being tested. In deciding upon the character of the tests to be made, and the extent to which they should be applied, one of the most important factors

[4] American Institute of Accountants, "Audits by Certified Public Accountants," New York, 1950, p. 23.

taken into consideration is the system of internal control. When evidence exists that the system is effective, the CPA properly concludes that the accounting records and supporting data have a higher degree of dependability than would otherwise be the case, and limits his testing accordingly. However, when his investigation shows that the system has points of weakness, he extends the scope of his testing. If the internal control is considered grossly inadequate or ineffective, he may feel compelled to review the entries in considerable detail before he can express an informed opinion on the financial statements.

Two comments are in order. According to this statement, if the accountant judges internal control to be good, he uses a small sample; if he judges it to be weak he uses a larger sample. This practice relegates sampling to the minor role of verifying a decision which has already been made. From the viewpoint of probability statistics this is putting the cart before the horse. Probability statistics, in the form of principles of sampling and estimation and inference, can be employed to obtain evidence which in itself can be used to appraise whether control is adequate or not. Furthermore the size of the sample can be based upon objective considerations, such as the variability of characteristics, the frequency of occurrence of errors, the magnitude of a difference, and the amount of sampling error that can be tolerated.

Actually the rule cited is not consistent with sampling principles—if internal control is very bad and the error rate is high in the several transactions it will require only a relatively small sample to detect this; on the other hand if the system of control is very good and errors are rare it will require a very large sample to detect one of these errors.

Finally we wish to emphasize the need for correcting certain notions about sampling which appear to be held in the fields of auditing and accounting. In the majority report of a subcommittee of the American Society of Mechanical Engineers, made in 1912, we find the following statement: [5]

> Persons having time and motion study in charge should possess that rare, intuitive human quality that causes the possessor to know when enough observations have been collected to form a sound working conclusion.

However, in a book on industrial internal auditing published about 40 years later (in 1951) we find the following statement: [6]

> This (the appropriate degree of test checking) is a subject which has received a considerable amount of treatment in public accounting articles and which is equally applicable to internal auditing. Unfortunately, however, no one has yet formulated any satisfactory substitute for the judgment of an experienced auditor, who through years of experience has gradually developed a "sense of touch" in this respect.

Today we do not have to rely upon an elusive something called "sense of touch," nor upon some "rare, intuitive human quality," in order to determine how large a sample of observations to

[5] C. B. Thompson, editor. *Scientific Management.* Harvard University Press, Cambridge, 1914, p. 168.

[6] W. A. Walker and W. R. Davies. *Industrial Internal Auditing.* McGraw-Hill, New York, 1951, p. 491.

take; we can use the principles of probability statistics applied to sample design. This does not mean, of course, that the experienced specialist cannot be of help in the design of a probability sample—he can often contribute useful knowledge about sources of variation, range of values, and certain difficulties involved in trying to sample the population.

PROBABILITY STATISTICS APPLIED TO INTERNAL CONTROL

A probability sample need not be limited to the periodic test check of the independent auditor; it can be applied to various types of internal control problems by accountants and controllers and statisticians employed by the business. This latter approach is much more effective because it is direct, immediate, and continuous. When continuous probability controls are placed on key operations, such as those associated with accounts receivable, payrolls, and inventory, trouble in these operations and errors made by key personnel are detected rather quickly so that they can be corrected at once; it is unnecessary to wait for an annual audit to reveal them.

Probability statistics is a "natural" for purposes of internal control. By means of sampling, estimation, experimental design, and inference based upon probability, this science provides the maximum amount of high quality data at a minimum cost, it provides ways of detecting many kinds of trouble, it provides a method of appraisal, and a method of improvement. Furthermore it can be applied to practically every major division and operation in the organization whether it is research and development, production, inspection, purchasing, sales, accounting, finance, or legal. Statistical quality control charts in production and sampling inspection in receiving are but two of many statistical methods which can be used for internal control purposes.

Consider the application of probability statistics to cost accounting, a form of internal control. Since so-called identical products are not identical but contain varying amounts of raw materials, require varying amounts of labor, and require varying amounts of machine time, it is clear that the unit costs of such products are not identical. This means that cost per article must be some kind of average cost. In order to obtain an average cost that is valid and meaningful it must be based upon sound methods of sampling, testing and estimating. The cost of a product, like a physical characteristic, cannot be safely predicted unless the process of production is under statistical control. For unless this control exists there is no stability to the behavior which we are trying to predict so that estimates are merely guesses and inferences are merely conjectures.

The cost accountant uses data derived from time and motion studies; these involve sampling and estimation and experimental design. He may run tests himself relative to unit costs; this also calls for the design of a test or experiment. He correlates characteristics; this calls for the use of regression analysis. He interprets differences and

variations; this calls for the use of statistical tests of significance and control limits.

The cost accountant deals with labor costs which are related to such statistical problems as learning curves, worker variability, sampling inspection, statistical control charts, error rates, and productivity analysis. He deals with material costs which are related to such statistical problems as dimension control, capability studies, receiving inspection, materials research, input-output analysis, vendor certification of statistical quality control, waste control, and container analysis.

AN EXAMPLE: THE CONFIRMATION OF ACCOUNTS RECEIVABLE

In order to illustrate current practice of sampling and interpretation of data obtained by the independent auditor in circularizing accounts receivable, the following description is quoted from a book on auditing: [7]

> In a few cases the whole customer list, including even those whose accounts show no balances at the year end, is circularized; but the testing principle is usually followed. This involves selecting the largest accounts that collectively will total say one-half of the dollar amount outstanding, and picking at random another ten per cent or more of the balance, making sure that samples are selected from each class of customers. . . . An answer cannot be expected from every request sent out; but if favorable replies are received from at least one-half (under the first method) or unfavorable replies from not more than one per cent (under either method), the propriety of the balance may be assumed.

In this situation neither the size of the sample nor the amount of the risk to be taken need be set in arbitrary terms; they can be based upon the application of probability sampling. The sampling of large accounts 100 per cent and the use of a stratification of the various classes of customers are sound sampling procedures; however the use of a ten per cent sample may be too large, or too small, depending upon what is being estimated, while the bias which is introduced by the nonrespondents to mail inquiries may be very serious. The recommended sampling practice in this latter case is to take a small random sample of the nonrespondents and to correct the bias in this way.

Nothing will be gained by using probability methods in part of the test and abandoning them in the later parts of the study; it is just as important to use probability methods to estimate characteristics and to interpret the data as it is to use probability sampling methods in the collection phase. Indeed the value of the latter is lost unless full advantage is taken of the former.

Consider a large public utility company which has 50,000 accounts receivable and about half with unpaid balances. In a book for certified public accountants it is asserted that a sample of a "few hundred" accounts will be adequate to test the adequacy of internal control. Whether a sample of this size is adequate or not depends upon what one is trying to do with the sample.

[7] Kohler, op. cit., p. 92.

AN EXAMPLE: THE AGING OF ACCOUNTS RECEIVABLE

A good example of the application of probability statistics to a problem in accounting is found in a recent paper by Trueblood and Cyert on the use of probability sampling for the purpose of aging the 100,000 accounts receivable in a large department store.[8]

"Aging" means that the total amount of accounts receivable is distributed according to the age of the account, a loss factor based upon past experience is applied to each age group, and these losses summed over all age groups in order to find the total amount uncollectible. The problem is to get an adequate estimate of this latter quantity.

The method used in this store was to draw a 15 per cent sample of the 500 trays containing 200 accounts each, for a total sample of 15,000 accounts. All accounts exceeding $2,500 were included. Then the accounts were "aged." After this was done the public accountant audited a ten per cent sample of these 15,000 accounts, or 1,500 accounts in order to determine whether they were aged properly.

Both of these procedures were modified. A random sample of 85 trays was drawn, and then every 10th account within each tray was selected to give a sample of 1,700. Accounts over $2,500 were still included 100 per cent. It was found that this greatly reduced sample would meet the specifications required by the accountants. The random sample of 1,700 appears to be about that given by the following specification for an unrestricted random sample: probability level of 95 per cent, two per cent error in the mean to be estimated, and a coefficient of variation of 0.41 for the accounts receivable distribution:

$$n = \left(\frac{2 \times .41}{.02}\right)^2 = 1681$$

The sampling error in the grand mean will be less than two per cent when the 100 per cent sample portion is combined with the sampled portion.

A sequential sampling plan was used to test the quality of the work done on the sample. The specifications set by the public accountants are as follows:

$p_1 = .03, p_2 = .08, \alpha = .10,$ and $\beta = .05.$

These values are of special interest because they show what risks the public accountant is willing to take in a problem of this type. Stated in words these specifications mean that in a problem of this kind the accountant is willing to take the risk of rejecting work with a three per cent error ten per cent of the time, and the risk of accepting work with an eight per cent error rate five per cent of the time.

This sampling plan results in a decision with an average sample size of 126, and even if the plan is applied to individual clerk control rather than for the purpose of general control, it still would require in general less sampling than the orthodox method of using a ten per cent sample. It should be noted

[8] R. M. Trueblood and R. M. Cyert, "Statistical sampling applied to aging of accounts receivable." *The Journal of Accountancy,* March, 1954, pp. 293–298.

that the sample size depends upon the quality of the work submitted by the clerks, and could be much less than 126.

Additional aspects of this problem which might warrant further investigation are the following: the possibility of using more stratification and reducing the sample size still further; this involves balancing costs since we can spend more on reducing the sample than the reduction is worth. The error introduced by using loss ratios not carefully determined may be much more than the error due to the sample of 1,700. Finally there is the problem of whether the public accountant needs to go into process control (that is, reaching a decision about the work of every clerk) or whether an overall test is sufficient.

AN EXAMPLE: THE CONTROL OF PAYROLLS

In a publication on auditing procedures issued by the American Institute of Accountants,[9] two cases are cited relative to fraud in payrolls, one in which the payroll clerk was in collusion with about 50 workers who paid him a commission on padded payroll amounts, the other in which a clerk took advantage of a rounding practice in connection with payrolls in order to pick up some petty cash. These examples suggest that sampling inspection techniques applied to payroll entries could be used to deter such practices, if not prevent them.

[9] American Institute of Accountants, "Papers on Auditing Procedure," New York, 1939, p. 108.

The following procedure might be used in connection with payrolls. After each payroll is prepared, the personnel are grouped on some significant basis such as salaries and wages, and within wages according to department, occupation, etc. Within each of these strata a small random sample is drawn. Then a careful examination is made of the basic payroll data for each of the names drawn: Time actually worked, rate of pay, premium pay if any, various deductions if any, leave if any, and other pertinent items. From these data estimates are made and inferences drawn.

It is recommended that in the beginning a straightforward stratified random sample be used in order to obtain estimates of the various characteristics. Later it may be possible to use a more sophisticated sampling plan, such as sequential sampling. In any event under a probability sampling plan, it would not be possible for collusion with 50 workers to remain undetected very long. Actually the very existence of such a current sampling audit system might deter a clerk from attempting to pad the payroll. Obviously such an audit system should be made independently of the payroll staff but they should be aware that such a current system of control is being used. Indeed it might be possible to adopt a system used by Alden's of Chicago in which large wall charts show the trend of the error rate performed by a given department or section. It is reported that some concerns are going even further and rewarding with bonuses or other types of compensation, workers or groups of workers who maintain a lower

rate of error (a higher level of quality) than some reasonable level.

So far as possible, sampling should be applied separately to the work of each key individual rather than to the aggregate of the work of several individuals. The reason is obvious: It allows better control over the individual and makes it possible to quickly find the source of error and take remedial steps to reduce or to eliminate it. On the other hand, where sampling is from a conglomerate of the work of many individuals it may be difficult to pinpoint trouble, so that any remedial work must be of the shotgun variety.

In this connection it is possible to determine how large a random sample is necessary in order to detect a condition of this kind. Assume that there are 1,000 workers on the payroll, that it is prepared weekly, and that the clerk has padded the wages of 50 workers. Let us answer this question: What is the probability that a random sample of 40 names taken independently every payday will detect this practice of padding by the end of three weeks? It is assumed that the practice is detected if one case is found in the sample. What we have is a random sample of 120, over three weeks, and a probability of occurrence of five per cent (50/1,000); hence the probability of getting at least one of these 50 names at the end of one week is $1-.95^{40}$ or .87, at the end of two weeks it is $1-.95^{80}$ or .98, while at the end of three weeks it is $1-.95^{120}$ or .994. Put another way this means that the probability of a complete miss is 1 in 8 for one week, 1 in 50 for two weeks, and 1 in 167 in three weeks.

If the sample is stratified by occupation, shift, size of wage, and other factors which might correlate with payroll padding, it would be possible to detect such practices with smaller samples.

Even in the case of rounding differences which were used as a means of petty stealing, it would be possible for probability sampling to detect this practice sooner or later. In the case cited the clerk would enter an actual amount of $25.62 as $25.63. Under the rounding rule he would raise this to $25.65 and collect this amount of money, but actually pay the worker $25.60, pocketing the nickel difference. Since this is a bias of five cents, a sample would detect this very quickly if the bias applied to many workers since such errors are not likely to occur very often by chance.

Assume that this practice applied to 100 workers out of a total of 5,000. How large a random sample of the payroll is required in order to insure a probability of 0.99 that at least one of these changes is detected? If we solve the equation $1-.98^n = 0.99$ for n we obtain 228 which is the size of the random sample required.

AN EXAMPLE: THE CONTROL OF INVENTORY

Consider the problem of maintaining control over the inventory of a grocery chain which operated over 800 stores.[10] A central warehouse controlled the inventory of each store by the retail method of charging the store with the selling price of goods shipped and of

[10] American Institute of Accountants, "Codification of statements on auditing procedure," 1951, pp. 35–37.

crediting it with sales. The method used was that of comparing the value of the inventory as revealed by a physical check with the value of the store's inventory account in the central office. The practice followed was that of using inventory crews to make three inventories a year in addition to the regular end-of-year inventory. "If a large shortage or overage developed, further investigation was made."

The difference between the book value and the physical inventory value is subject to four major sources of error: the book figure itself may be in error, there may be a bias in taking the physical inventory, there may be an error in estimating the amount of store loss due to spoilage and damage and the like, and there may be stealing or otherwise a diverting of merchandise. Clearly the first three sources have to be carefully controlled if the fourth source is to be detected, or otherwise this source will be confounded with other sources of error. This is particularly true if the order of magnitude of this last factor is about the same order of magnitude as that of the other sources of error. In other words, whether we get adequate control or not depends upon whether we eliminate the first three sources of error.

In this chain over 2,400 physical inventories are made annually in addition to the end-of-year inventory. It seems likely that as good or even better control could be obtained by using probability sampling methods; one such method might be the following:

1. Divide the stores into 40 groups, with the largest stores in one group, the suspect stores if any in another group, and all other stores stratified by size of inventory.
2. Draw five stores at random from each of the 38 groups; take all stores in the large and suspect groups.
3. Do a careful physical inventory of these sample stores.
4. Estimate total inventory value of each group, and compute the difference between this value and the book value of that group; do the same for the inventory of all stores.
5. Test whether the difference for each group, as well as for the total, is outside tolerance limits set on a probability basis.
6. If a group value is within the limits accept the group, if not within the limits do a physical inventory on the remaining 15 stores in the group.
7. Keep control charts on each group of stores as well as on all stores as a group.
8. Make this test five times per year.
9. Samples can be drawn so that every store is selected at least once per year.

The foregoing plan requires as a basis for its success, a very careful control over the first three sources of error given above. Due to the fact that a smaller total number of physical inventories can be taken each year under a sampling system, a much more careful job can be done on each inventory. Furthermore, it is necessary to keep careful records of the losses due to damage and spoilage so that a correct adjustment can be made for these elements. Finally the book figures must be carefully checked. When this is done, the error in the difference found is due primarily to two factors—the random variation due to sampling which can be calculated, and the bias due to the fourth factor which we are trying to detect. Under these conditions when a

sample point (derived from a random sample of five stores) falls outside the probability limits it indicates that the fourth factor is operating.

The proposed sampling plan calls for about 1,000 inventories per year instead of 2,400. Stratified random sampling makes it possible to exert control through groups as well as to obtain adequate estimates of population values with the minimum amount of sampling. In this method not only is every store subject to control through the group five times per year on a sampling basis, but is subject to direct control at least once per year. All this is in addition to the end-of-year physical inventory.

In an actual case the independent auditors took a judgment sample of eight stores as a basis for testing the inventorying of 200 variety stores. If a random sample of eight stores had been selected from the 200, what is the probability that a wrong decision would be made if ten per cent of the stores had a defective internal control system? If the percent was 20 per cent? 30 per cent? The results are summarized below:

Percent of stores with poor control	*Number with poor control*	*Probability that a sample of 8 will lead to approval*
10	20	.43
20	40	.13
30	60	.06

What this shows is that a large risk is assumed in trying to reach a decision on the basis of a random sample of only eight stores. For example, if 20 of these 200 stores have defective systems of control about 43 per cent of the time on the basis of a random sample of eight stores, the conclusion will be reached that control within the 200 stores is adequate.

SOME EXAMPLES: THE CONTROL OF COSTS

Cost accounting is concerned with both quality control and with quantity control. It is concerned with quality control because its aim is to obtain a high quality product at the lowest cost, and it is concerned with quantity control in order to protect the assets of the company and to minimize the consumption of production factors for a given amount of output. Since probability statistics can be applied to both quality and quantity control, it has a wide field of application in the field of cost accounting. We shall cite three examples.

In setting standards for materials use, the problem arises of determining the average amount of material per unit of product. The amount of nylon yarn per pair of nylon hose, already cited, is a case in point. In a book of standard costs [11] this problem is illustrated by reference to the amount of paint required per body of a toy express wagon. It was suggested that an operator be instructed to use a spray gun for a specified length of time on each body according to instructions given by the time study man. A run of 30 bodies is made and the total amount

[11] S. B. Henrici, *Standard Costs for Manufacturing*. McGraw-Hill, New York, 1953. Second edition, pp. 95–128.

of paint calculated. From this figure the average amount of paint per wagon is obtained. It is stated that conditions must be at the best attainable level in order for this test to be valid.

In this type of problem where we are trying to determine the amount of material required per unit of product, it may be necessary to make these quickie tests. However, if one wants high quality information, wants to determine the best obtainable levels, wants to measure the influence of major variables, wants to maximize the information at the minimum cost, it is necessary to design an experiment and interpret the data on a probability basis.

Consider another problem where the cost accountant wants to find out the amount of compressed air required to produce an acceptable finish in a minimum time by using shot blast for cleaning castings. In this problem the amount of air used is a function of several variables including air pressure, the size of the shot used, the degree of finish desired, the type of nozzle used, and the location of the nozzle relative to the work. In order to determine this relationship so that an optimum situation can be maintained, it is necessary to design an experiment and to apply the method of partial regression to the data.

Another cost problem where probability statistics can be used to good advantage is that of estimating and controlling various types of wastes and losses. Noble has described, for example, how statistical quality control methods can be used for controlling waste in the paper manufacturing industry.[12] The specific problem was one in which the cost accountant was asked to set up a system of controlling waste in a certain department converting rolls of paper into sheets. The data recorded for three shifts for a period of ten days were as follows:

	Days										
Shift	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*Sum*
A	89	112	121	91	75	86	123	98	96	97	988
B	99	108	106	117	79	105	106	100	83	114	1,017
C	115	132	103	98	81	93	105	114	87	124	1,052
Sum	303	352	330	306	235	284	334	312	266	335	3,057

An analysis of variance shows that these data indicate a state of control existed.

What these data show is that the waste produced in this department is a variable and not a constant, and if it is controlled, cost, like production, will fluctuate about an average value in an approximate random fashion, the extent of which can be calculated.

SUMMARY

General accountants, cost accountants, independent auditors, and internal auditors are faced with numerous prob-

[12] Carl Noble, "Statistical Cost Control in the Paper Industry." *Industrial Quality Control*, Vol. IX, No. 6, May 1953; p. 42.

lems which can be solved by means of probability statistics. This science provides a specific and effective *modus operandi* for improvement: first it provides a way of determining the current situation, second it provides a way of determining materials and machine and process capabilities, third it provides a way of controlling many operations, and fourth it provides a method of appraising various materials and methods and processes. Two broad implications may be mentioned. Accounting, auditing, and internal control methods generally need to be reconsidered in view of this science. Finally in order to obtain the maximum benefit from probability statistics it is necessary to have much more cooperation than now exists between accountants and auditors on the one hand and probability statisticians on the other.

✦✦✦✦✦✦✦✦✦✦✦✦ CAPITAL BUDGETING AND GAME THEORY

EDWARD G. BENNION

The purpose of this article is to do two things: (1) to suggest a more rational approach to capital budgeting, which is a perennial and imperfectly solved problem for business, and (2) to test the applicability of game theory to the kind of decisions which are involved in capital budgeting.

In a sense, this is singling out a particular problem and a particular technique. However, the problem happens to be one of the most important and least clarified of the top-level issues faced by businessmen, just as the technique happens to be one of the most intriguing and least understood of the statistical devices which have recently been presented to businessmen as aids to top-level decision making. In combination, they offer an unusual opportunity to push progress ahead in an area where it is needed and at the same time put some realism into a methodology whose value may be overestimated at its present stage of development.

ROLE OF FORECASTS

The whole subject of capital budgeting is, of course, too big and complicated to be critically examined within the scope of one brief article. But perhaps some new light can be thrown on one important problem aspect: the relationship that should obtain between the budget and the economic forecast.

It hardly seems necessary to prove that economic forecasts play a significant—sometimes an almost determining—role in shaping the businessman's investment decisions. The traditional explanation is simple:

> In order to decide how much a company should invest or what kinds of assets it should acquire, we need a sales forecast

Harvard Business Review, *November–December 1956, 34:6, 115–123.*

for the firm to establish its anticipated level of activity.

But the firm's sales forecast cannot be made without some estimate of what the industry is going to do. And the industry's sales forecast in turn depends in large measure on the predicted level of activity in the economy as a whole.

Q.E.D.—the capital budget of any individual firm has a unique and important relation to the general economic forecast.

UNRELIABLE GUIDE

If it is obvious that forecasts are necessary, it is still more obvious that they are likely to be unreliable:

It is impossible to make an economic forecast in which full confidence can be placed. No matter what refinements of techniques are employed there still remain at least some *exogenous variables*—i.e., variables, such as defense expenditures, the error limits of whose predicted values cannot be scientifically measured.

It is thus not even possible to say with certainty how likely our forecast is to be right. We may be brash enough to label a forecast as "most probable," but this implies an ability on our part to pin an approximate probability coefficient on a forecast: 1.0 if it is a virtual certainty, 0.0 if it is next to an impossibility, or some other coefficient between these extremes. But, again, since we have no precise way of measuring the probability of our exogenous variables behaving as we assume them to do, there is no assurance that the *estimated* probability coefficient for our forecast is anything like 100% correct.

In spite of such drawbacks, businessmen are willing to pay for having general economic forecasts made, and to use them in deciding among alternative investment opportunities for capital funds. For example, the more certain is prosperity, the wiser it will usually be to invest in new plant and equipment, whereas the more certain is depression or recession, the safer it looks to invest in government bonds or other securities. In other words, the businessman uses economic forecasts to assess the relative advantages of investing in fixed or liquid assets, in the light of the expected business-cycle phase.

COMMON ERROR

At this point, the businessman stands before us, his economic forecast in one hand, his proposed investment alternatives in the other. His next step is the one where he is most apt to go wrong. When some one phase of the business cycle is forecast as "most probable," it is likely to look logical to him to go ahead and put his funds into whichever investment alternative maximizes profit in the phase expected.

Looking at the situation superficially, this step appears to be quite sensible. But actually a businessman armed with only a single most probable forecast is in no position to make a wise investment decision—*unless* his forecast is 100% correct, and this, as we have seen, is an impossibility.

A NEW APPROACH

In the following pages, a more rational way to use an economic forecast is suggested. Furthermore, adoption of the method proposed here permits the businessman to learn the answer to another question over which he probably has spent some sleepless nights if he has ever known responsibility for making a decision on the capital budget.

Just how far off can the forecast be before it leads to a "wrong" investment decision?

Because the fundamentals of this new approach are most easily grasped if a specific problem is attacked, let us see how it can be applied in concrete cases. We shall look first at a simplified hypothetical case, and then at a case based on actual experience (slightly disguised). For the sake of the clearest possible focus on the problems involved, no explicit reference will be made to the role of game theory while we are working out their solution. Following their presentation, however, we will meet the theory head-on and discover in the process that we have already drawn from it just about as much as is possible.

SIMPLIFIED CASE

This first case, although hypothetical, is not unrealistic. Further, it has the advantage of reducing the problem and method of solution to the simplest possible proportions.

ALTERNATIVE INVESTMENTS

The specific issue of whether to invest in plant or securities is a good one for illustrative purposes because it can be defined so sharply. Suppose we have even more exact information than most businessmen generally assemble before exercising their judgment to reach an investment decision. Under these circumstances we might know that:

The most probable forecast is for a recession.

In recession, investment in plant will yield 1% as compared with a 4% yield for securities.

In prosperity, plant will yield 17%, while securities will yield 5%.

Placing these data in diagram form, we get the following 2×2 "matrix":

		Cycle-Phase Alternatives	
		Recession	*Prosperity*
Management Investment Alternatives	Securities	4%	5%
	Plant	1%	17%

Under this condition no businessman worth his salt is going to want to settle for securities—but how can he justify any other course, given his forecast of a probable recession?

MORE DATA NEEDED

To begin with, our businessman needs to recognize that the data so far placed at his disposal, rather than limiting his choice, do not provide the basis for a decision at all. Two further questions first require an answer:

(1) How probable is the "most probable" forecast? To answer this, the forecast needs to be completed by assigning a *probability coefficient* to each cycle phase considered.

(2) How probable does a recession have to be before the earnings prospects of the more conservative choice look just as attractive as the returns available from adopting a bolder course of action? In other words, what are the *indifference*

probabilities of recession and prosperity, given the rate of return each will yield?

PROBABILITY COEFFICIENTS

Establishing probability coefficients on the economic forecast is a job we can relinquish, more than willingly, to the company economist. We are not concerned here with what kind of crystal ball he gazes into, but rather with how top management uses his findings, whatever they may be. So, in order to get on with our problem, let us simply suppose that our forecaster thinks the chances of recession are 6 out of 10 and so has assigned a probability coefficient of 0.6 to his predictions for a recession (which automatically means 0.4 for prosperity).

INDIFFERENCE PROBABILITIES

The handling of indifference probabilities is not going to be quite so simple, but it can be done readily enough by anyone who can recall his high school course in algebra (he does not have to be blessed with "total recall," either):

Suppose we say, in elementary algebraic terms, that the recession probability coefficient = R, and the prosperity probability coefficient = P. In that event, we know from our matrix figures that over any period:

I: $4R + 5P$ = the return on securities
II: $1R + 17P$ = the return on plant

From this, it is clear that the return on securities will be the same as the return on plant when:

III: $4R + 5P = 1R + 17P$

Solving this last equation for R in terms of P, we get:

IV: $R = 4P$

Since the sum of the probability coefficients $(R + P)$ has to equal 1.0, we can say that $R = (1 - P)$ and substitute $(1 - P)$ for R in IV:

V: $1 - P = 4P$, or
VI: $P = 0.2$, and $R = 0.8$

This merely means that if the probabilities of a recession and prosperity are 0.8 and 0.2 respectively, then the chances are that the company will be just as well off investing in securities as in plant, and vice versa. In other words, it appears to be a matter of *indifference* which alternative is chosen.

ASSEMBLING THE DATA

For the sake of convenience, let us reassemble all our information in the compact easy-to-read form of a matrix:

		Cycle-Phase Alternatives	
		Recession	*Prosperity*
Management Investment Alternatives	Securities	4%	5%
	Plant	1%	17%
Indifference probabilities:		$R = 0.8$	$P = 0.2$
Forecasted probabilities:		$\hat{R} = 0.6$	$\hat{P} = 0.4$

Here, at last, our businessman has all the information he needs to decide what course of action maximizes his chances for success. So long as the forecasted probability coefficient for a recession is not equal to or greater than the indifference probability coefficient for this phase of the business cycle, the businessman can know that he is not making an *avoidable* mistake by playing for high stakes and building a plant. *The alternative to choose is the one that has a higher forecasted probability than indifference probability.*

ADVANTAGE GAINED

The little technique outlined above thus does two things:

(1) It makes clear that the best paying investment alternative in the most probable situation is not necessarily the alternative that management should choose.

(2) With indifference probabilities, it is possible for us to see what margin of error is permissible in any estimated probabilities before these estimates result in an erroneous decision.

ACTUAL CASE

With this much understanding of the 2×2 matrix, we are now in a position to apply indifference probabilities to our actual but more complicated case:

An integrated petroleum company anticipates the need for a refinery in Country A, has determined that Alpha City is the best location, but is uncertain as to the appropriate size.

Operating at approximate capacity, internal economies of scale exist up to a refinery size of R barrels per day (B/D).

However, once sales exceed Z barrels per day (with $Z<R$), further economies could best be effected by building a second refinery elsewhere. This puts a ceiling of Z barrels per day on the Alpha City unit.[1]

Whatever size refinery is built, it can be completed in 1960. It is also agreed that depreciation and obsolescence will make the refinery valueless by 1974.

As an aid in determining the size required in 1960, it is known that consumption growth is highly correlated with industrial output.

Unfortunately, there is less than perfect unanimity as to the expected growth rate of industrial output between the present and 1960. The economics department has forecasted a rate of 2.5%, the foreign government officially estimates a rate of 5%, and the company top management wonders what would happen if the growth rate turned out to be 7.5%.

Careful analysis leads to the conclusion that a growth rate of 2.5% requires a refinery of X barrels per day capacity; that a growth rate of 5% necessitates a refinery of Y barrels per day capacity; and that a growth rate of 7.5% requires a refinery of Z barrels per day capacity, this last being our previously established ceiling size.

Pending further study, all agree to work on the assumption of a zero growth rate after 1960.

To evaluate the three alternative refineries, anticipated integrated income (covering refining, marketing, produc-

[1] This ceiling decision, it might be noted, involves the solution to a problem to which linear programing conceivably might aptly be applied. Taking this solution as given obviously does not mean it is necessarily easy to come by. See Alexander Henderson and Robert Schlaifer, "Mathematical Programing: Better Information for Better Decision Making," HBR May–June 1954, p. 73.

ing, and transportation) will be computed for each facility under each growth rate, and the per cent return on integrated investment will then be calculated and compared.[2]

But filling in the matrix does more than verify our common-sense conclusions. It also equips us to see *how much* better one refinery is than another under each possible condition.

		Pre-1960 Growth Rate Alternatives		
		Low 2.5%	*Moderate* 5.0%	*High* 7.5%
Refinery Investment Alternatives	Z B/D	2.0%	7.3%	12.6%
	Y B/D	3.7%	11.0%	11.0%
	X B/D	8.8%	8.8%	8.8%
Indifference probabilities:		$L = 0.301$	$M = 0.114$	$H = 0.585$
Forecasted probabilities:		$\hat{L} = 0.333$	$\hat{M} = 0.333$	$\hat{H} = 0.333$

MATRIX AND PROBABILITIES

With three sizes of refineries to consider, and three growth rates, we will get a 3×3 matrix on this problem. The figures on the diagram, representing return on integrated investment, are more or less what common sense tells us to expect. For example:

The small X B/D refinery shows the highest rate of return if the growth rate is a low 2.5%, bringing in an 8.8% return against only 2% for the large Z B/D unit with its much higher cost of investment.

On the other hand, if the growth rate should reach a high of 7.5%, the large Z B/D refinery can return an average of 12.6%, while the small X B/D facility with its limited output is tied to its 8.8% ceiling yield.

[2] While the matrix figures of this case are based on careful engineering estimates, this venture is still in an experimental stage and hence does not constitute a part of the budget procedure of Standard Oil Company (New Jersey), with which I am associated.

The indifference probabilities here were calculated by just the same algebraic procedures as were followed in our previous example. The forecasted probabilities simply reflect the fact that no one in the company could decide which of the three forecasts was most likely, and therefore each was treated as equally "valid" (i.e., chances of 1 out of 3, or $0.33\frac{1}{3}$).

THE SOLUTION

A quick look at our diagram now reveals that the extra-large Z B/D refinery should be ruled out, since its return will be greater only under a 7.5% growth rate, and the real probability for a 7.5% growth rate is too small to justify considering that alternative. (Remember that an alternative cannot be chosen unless its forecasted or estimated true probability is equal to or above the indifference probability.)

On the other hand, both the X B/D and Y B/D refineries are still in the running. So, with two possibilities still remaining, new indifference probability calculations are needed in order that we may choose between them:

Suppose (as before) we use the letter L to represent the indifference probability for the low 2.5% growth rate; M for the moderate 5% rate; and H for the high 7.5% figure. In this event we read off the matrix that:

I: $3.7L + 11.0M + 11.0H =$ the return on the Y B/D refinery

II: $8.8L + 8.8M + 8.2H =$ the return on the X B/D refinery

From this it is clear that the return on Y B/D will be the same as the return on X B/D when:

III: $3.7L + 11.0M + 11.0H = 8.8L + 8.8M + 8.8H$

Solving this equation in terms of L we get:

IV: $L = 0.43\ (M + H)$

This means that if the estimated true probability of a 2.5% growth rate is greater than 0.43 of the combined estimated true probabilities of the 5% and the 7.5% growth rates, the X B/D refinery is a better bet than the Y B/D refinery. This has to be true because the X B/D refinery is the best-paying alternative, given the 2.5% growth rate.

Since the estimated true probability for the 2.5% growth rate is actually 0.33, which is slightly greater than $0.43 \times (0.33 + 0.33)$, we would conclude that the X B/D refinery is a little better bet than the Y B/D refinery, with the Z B/D refinery showing a very poor third.

A CHANGED ASSUMPTION

Now that we have reached an answer to the problem as originally stated, let us (realistically if provokingly) proceed to alter some of our assumptions, and see just what this will do to our choice:

Suppose that top management, having injected the 7.5% growth rate into the original problem for comparative purposes, concludes that the probability of a 7.5% growth rate is really nil, and that the probabilities of the 2.5% and 5% growth rates are each 0.5. The indifference equation then becomes:

$$3.7L + 11.0M = 8.8L + 8.8M, \text{ or}$$
$$L = 0.43M$$

Since the estimated true probability of 0.50 for the 2.5% growth rate is, in this instance, a great deal bigger than 0.43×0.50, we would conclude that the X B/D refinery is a lot better bet than the Y B/D refinery, and that no one in his right mind would even consider building a Z B/D unit.

Thus we come to the same general conclusion as before, only a bit more cocky, as a result of writing off the 7.5% growth rate and distributing its former probability in such a way as to make the 2.5% and 5% growth rates equally probable.

A RADICAL REVISION

It is clear, however, that the above conclusion is suspect unless we expect no economic growth in Country A after 1960. If we do expect further growth, the X B/D refinery loses much of its $-sign allure. It will not be large enough

to take advantage of Country A's expanding economy and so can never return any more than 8.8% on investment.

In contrast, the Z B/D refinery will show an increasing rate of return as A's expanding market permits it to produce more and more per year, perhaps ultimately reaching its capacity. Thus, instead of spurning the Z B/D refinery (as in our last example), we must acknowledge its potential attractiveness—provided A's economy does not get stalled after 1960, as was previously assumed.

With this possibility in mind, let us make some alterations in our problem and see what we should do. There are two new conditions:

(1) After 1960, it is now agreed, the growth rate of Country A will be a steady 2.5% each year.

(2) It would be possible to build an X B/D or Y B/D refinery that would be expansible to Z B/D; such units would cost more than nonexpansible facilities, but less than two separate refineries with a combined Z B/D capacity.

At this point our real problem becomes one of deciding whether to build a Z B/D refinery or an X B/D or a Y B/D refinery expansible to Z B/D. It may be helpful, first, to consider how the figures in this new matrix should differ from those presented earlier:

Most of the figures are higher than before, reflecting the fact that average earnings are increased by higher sales toward the end of the productive life of each unit.

To a limited extent, the higher investment costs of the two expansible units operate as a drag on their earnings. Thus three figures happen to be lower than before, and the expansible refineries have a lower maximum return than is possible with a Z B/D unit.

On the whole, the figures in the matrix tend to be squeezed closer together; i.e., they no longer range between such wide extremes.

Now our revised matrix reads like this:

		Pre-1960 Growth Rate Alternatives		
		Low 2.5%	*Moderate* 5.0%	*High* 7.5%
Refinery Investment Alternatives	Z B/D	4.6%	9.9%	12.6%
	Expansible Y B/D	7.2%	10.5%	10.9%
	Expansible X B/D	8.0%	9.7%	10.4%
Indifference probabilities:		L = 0.388	M = 0.014	H = 0.598
Forecasted probabilities:		$\hat{L} = 0.333$	$\hat{M} = 0.333$	$\hat{H} = 0.333$

Again, we work our algebraic equations to find the indifference probabilities, while the forecasted probabilities result, as before, from assigning equal weight to each forecast.

In this instance, the expansible Y B/D refinery is a shoo-in. This can be intuitively seen by recognizing that the Y B/D refinery is the best paying one given the 5% growth rate, and this growth rate is the only one with an *indifference* probability lower than its *estimated true* probability. By somewhat similar reasoning, the Z B/D refinery is distinctly the worst choice.

The new matrix also reveals another significant conclusion. One does not have to be a mathematician to perceive, just from inspection, that the cost of a poor decision here is a good deal lower than it was for our earlier versions of this problem. The result, of course, flows from the fact that the differences in the row and column vectors have been greatly narrowed. The *absolute* cost of a mistake is by no means insignificant, but *relatively* it is much less than in the previous matrix. This piece of information is in itself of considerable value. At a minimum, it will help the budget-maker to do less agonized tossing in his bed.

MORE TINKERING

Just for fun, let us tinker with our problem once more before dropping it, and again assume that management rejects as of nil probability the growth rate of 7.5% between the present and 1960, giving the 2.5% and 5% growth rates equal probabilities (i.e., chances of 1 out of 2, or 0.5):

This eliminates the top row and right column of the 3 × 3 matrix, leaving it 2 × 2.

The indifference probabilities equation for the X B/D and Y B/D refineries then becomes:

$$8.0L + 9.7M = 7.2L + 10.5M,$$
$$\text{or } L = M$$

This means that the *indifference* probabilities for both L and M are 0.5.

These, however, are also the values for the *estimated real* probabilities for the 2.5% and 5% growth rates.

Consequently, we have here the unusual case in which the X B/D and Y B/D refineries are equally good bets, with the Z B/D refinery being no bet at all.

No matter which way we look at the problem, therefore, the Z B/D refinery is the poorest choice. But, under one probability assessment, the expansible Y B/D refinery is a better choice than the expansible X B/D refinery; under the other, the expansible X and Y B/D refineries are toss-ups. This conclusion is, of course, decidedly different from that reached for the previous matrix, in which the influence of post-1960 growth was ignored.

OTHER BUDGET QUESTIONS

So far we have managed to explore only one small corner of the capital budget domain. Let us look at some further problem areas.

PROBLEM OF TIMING

Our new technique can also be put to work on a timing problem. Since all but one of the figures in the following matrix have been chosen somewhat ar-

bitrarily (although the choices can easily be defended), they call for no explanation. The only exception is the 9.6% prosperity return for the Y B/D refinery which appears in the lower right-hand corner; this is the *weighted average* return from this investment under the three possible rates of growth, assuming equal probability for each.

It would be a tedious repetition of now familiar principles to attempt to bleed this matrix dry. Let us content ourselves, therefore, with just one reasonable (and relatively simple) interpretation:

Since depression is comparable to the 1937–1938 decline in this country, we might well reject this as being of nil probability in the period between now and 1960.

Should we do so, the left column of the 3 × 3 matrix would be eliminated, as would the top row, since government bonds would not be a logical investment except under depressed conditions.

In the remaining 2 × 2 matrix, the indifference equation for other securities and our Y B/D refinery then becomes:

$$5.0R + 3.0P = 3.0R + 9.6P, \text{ or } P = 0.3R$$

In other words, unless we think the true probability of a recession is something more than twice as great as that of prosperity, the construction of the Y B/D refinery ought not to be deferred.

INVESTMENT PRIORITIES

Whenever more than one investment alternative is available, there arises the problem of assigning an order of priority among them. To assess and compare each possible project, the method followed in the previous problem can be used to advantage again:

(1) Using a 2 × 2 matrix, calculate indifference probabilities for investing in securities and in plant (or other assets to be used in the company's own business).

(2) Repeat this process for each contemplated internal use of company funds. The top row will be the same in all of these matrices, but the bottom rows will not in general be the same. Consequently, the indifference probabilities for the numerous matrices may vary widely. Any company project with an indifference probability coefficient for prosperity that is lower than the estimated true prosperity probability coefficient is a good bet.

		Pre-1960 Cycle-Phase Alternatives		
		Depression	*Recession*	*Prosperity*
Management Investment Alternatives	Government Bonds	3.5%	3.0%	2.5%
	Other Securities	2.0%	5.0%	3.0%
	Y B/D Refinery	−2.0%	3.0%	9.6%
Indifference Probabilities:		D= 0.43	R = 0.24	P = 0.33

(3) Instead of arraying the projects in order of descending return *for the most probable cycle phase* (which would incur all the defects already shown to exist in tying investment decisions to a single most probable forecast), array them in order of descending *weighted average* return—weighted according to the estimated probabilities of recession and prosperity or of different rates of growth.

In general, this procedure will *not* result in the same priority order for projects as the method commonly employed, but it is a better method of evaluating all the alternative uses for funds. This is because it avoids the frequently fatal mistake of betting on whatever venture seems to look most profitable, given only a single most probable forecast.

If desirable projects turn out to be more numerous than company resources can finance, the management must then decide whether it wants to borrow or not. If external financing should be ruled out, the marginal project must be the one with the lowest weighted average return which just exhausts available funds. On the other hand, if all desirable projects do not exhaust available funds, the marginal project is the one whose indifference coefficient for prosperity is just equal to the estimated true prosperity coefficient. The excess funds should be temporarily invested in securities.

ALMOST A GAME

Now, finally, we are ready to have that initial promise redeemed—i.e., that the role of game theory would be explained and be evaluated. Actually, as anyone who has met this theory before will recognize, it has already been introduced! All our matrices have been "games," although, in playing some of them, we have had to construe a few of the rules pretty loosely.

The two players in most of our games have been the businessman and the business cycle. Each has had either two or three "strategies." For the former there have been different types of investment alternatives; for the latter there have been different cycle phases. Indeed, the indifference probabilities calculated by the businessman for depression, recession, and prosperity have an exact parallel in game theory. Those probabilities constitute what would be known as the "business cycle probabilities"—namely, the percentage of the time that the business cycle should provide each of its phases in a random manner to hold the businessman's gains down to a minimum.

Can it be said, then, that our method for deciding on the capital budget marks an extension of game theory concepts to the field of business and economics? In the strictest sense, the answer must be *no*. Ours is not a rigorous game—it does not meet all the conditions requisite for such a game:

> In game theory proper, the opposing players are assumed to be completely selfish and intelligent. Charity and stupidity are unknown to either. Clearly the business cycle, however malevolent it may sometimes seem, does not meet these requirements. It is as impersonal as nature. In fact, what we really are doing in problems like ours is playing games with nature. Thus the indifference probabilities in our last matrix are actually *nature's*

probabilities. They tell us that, if nature were malevolent, it could minimize its "losses" to the businessman by providing depression, recession, and prosperity, in a random manner, 43%, 24%, and 33% of the time, respectively.

If the businessman were confronting an opponent who could maximize gains and minimize losses by a deliberate choice of strategies, then there would be additional calculations to make and prohibitions to observe. For example, in order to keep a selfish and intelligent antagonist from guessing what he might do and benefiting by the knowledge, the businessman might have to figure out several strategies for himself and then use them randomly. Thus, again in our last matrix, the *businessman's* odds are such that he should invest in government bonds, other securities, and the Y B/D refinery, in a random manner, 90%, 2%, and 8% of the time, respectively. Otherwise, faced with a malevolent nature, he would fail to maximize his gains.

It takes some stretching to make a choice of strategies out of a range of possibilities, yet a range of possibilities is all we can get out of nature (as contrasted with a willful opponent); and in the case of some business problems we cannot even get that. Moreover, nature, alias the business cycle, may have some strategies on the matrix that no sensible antagonist would use at all because under all conditions other strategies would give him higher gains or lower payoffs.

Consequently, our games with nature are not of the "purer," more rigorous type. Our version represents a departure by virtue of recognizing four additional facts: (1) nature is not malevolent; (2) the odds of a malevolent nature are really the indifference probabilities of the businessman with respect to his alternative courses of action; (3) any time the estimated odds of nature's strategies differ from the businessman's indifference odds, there is a best strategy for the businessman; and (4) this best strategy, as well as the degree of its "bestness," depends on the relationship between the estimated and indifference odds.

However, any readers who are interested in further pursuing the rules of game theory proper can do so handily by consulting J. D. Williams, *The Compleat Strategyst* (New York, McGraw-Hill Book Company, Inc., 1954). Anyone who can add and subtract can follow this pleasant and often humorous exposition, whereas most other books on the subject call for more advanced mathematical learning.[3]

OTHER ECONOMIC "GAMES"

If capital budgeting can only borrow from game theory but not take it over in its entirety, what about any other business applications? It should be possible to find in the businessman's competitive world a variety of situations that resemble orthodox games—i.e., where the opponents are not noted for their charity toward each other.

The existence of many such parallels is obvious, but unfortunately game theory in its present state of development is not far enough advanced to handle

[3] John von Neumann and Oscar Morgenstern, *Theory of Games and Economic Behavior* (Princeton, Princeton University Press, 1944); J. C. C. McKinsey, *Introduction to Theory of Games* (New York, McGraw-Hill Book Company, Inc., 1952); David Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions* (New York, John Wiley & Sons, Inc., 1954).

most of them. (Originated by von Neumann, the theory first achieved a wide audience when he and Morgenstern published their book in 1944.[4]) Thus, game theory still does not deal effectively with situations where there are more than two players or where the loser's losses and the winner's gains do not cancel out. For example:

The most common realistic game cited in economic literature is a duopolistic (two-seller) situation in which each of the duopolists has alternative strategies and seeks the strategy that will maximize his profits.[5] This may be a realistic example, but it is certainly one of limited existence. The businessman may not have a large number of competitors, but he usually has at least several. However, to consider several competitors plunges us into games involving more than two players, and here the theory as it now stands leaves much to be desired.

Another possible realistic game on the two-person level is where the opponents are the businessman and the trade union. But this sort of game is likely to be one in which the solution may harm or benefit both players, or harm one player more than it benefits the other. This throws us into games with a non-zero-sum payoff, where the theory again leaves much to be desired.

To say that game theory, in its more rigorous sense, still has no significant business applications does not of course mean that claims for its *potential* have been exaggerated. The day of orthodox game theory may well be on its way, just as the day of linear programing has already arrived in some measure.[6] Meanwhile, businessmen may wish to acquaint themselves with the theory and be on the watch for any practical uses it may have.

CONCLUSION

To summarize briefly, we have seen that forecasting can result in a negative contribution to capital budget decisions unless it goes further than merely providing a single most probable prediction. Without an estimated probability coefficient for the forecast, plus knowledge of the payoffs for the company's alternative investments and calculation of indifference probabilities, the best decision on the capital budget cannot be reached.

Even with these aids the best decision cannot be known for certain, but the margin of error may be substantially reduced, and the businessman can tell just how far off his forecast may be before it leads him to the wrong decision. It is in assessing this margin of error, along with the necessarily quantitative statement of alternative payoffs, that some of the concepts of game theory make their particular contribution to the problem.

[4] See footnote 3.

[5] See L. Hurwicz, "The Theory of Economic Behavior," in George J. Stigler and Kenneth E. Boulding (editors), *A.E.A. Readings in Price Theory* (Chicago, Richard D. Irwin, Inc., 1952), Vol. VI.

[6] See Alexander Henderson and Robert Schlaifer, op. cit. (footnote 1).

✦✦✦✦✦✦✦✦✦✦✦✦✦ THE PLACE OF SCIENTIFIC TECHNIQUES IN MERGERS AND ACQUISITIONS

ROGER R. CRANE

In considering a merger or acquisition, management faces one of its most difficult decisions. There is much potential risk involved because of the long-term implications of such a move. An ill-advised acquisition can be the turning point in a company's history—from profit to loss, from prestige to failure. On the other hand, a carefully planned acquisition can lead to success for both parties, open entirely new product areas, and provide a basis for a good profit and for healthy growth opportunities.

In recent years, many companies have chosen the acquisition route for expansion and have been faced with the difficult decision of selecting a company or companies with which to merge. Naturally, many different methods are used for determining the best candidate for acquisition. However, such planning generally includes a great deal of detailed analysis. Volumes of information may be gathered concerning the prospect's plant, its personnel, its products, and many other aspects of its profit potential. All available financial data are usually gathered and interpreted. This analysis often must be carried out under considerable secrecy, particularly in the event that competitors are interested in the same companies.

However, when the decisions are finally made, frequently they seem to be based largely on politics or on "feelings" about certain companies, rather than on sound, logical considerations. The detailed analysis may not be used adequately in the final decision at the board of directors level because of the natural difficulty of digesting the volume of paper and untangling the many interlaced decision elements in relatively infrequent and brief board meetings.

A certain amount of initial screening of merger candidates can be accomplished quite readily once basic guidelines have been established. Some companies which are obviously unsuitable can be immediately discarded without further consideration. There normally remains a residue of three or four almost equivalent companies, some strong in one desirable aspect and weak in others, and vice versa. Accordingly, the difficult problem of weighing and balancing the different factors to make the best decision is created.

Although the comparison of companies according to these factors can be done intuitively through the exercise of judgment based on experience, the greater the number of factors involved, the more confusing the situation becomes. As the confusion increases, the

The Controller, *July 1961, 326–329, 340, 342.*

possibility of a serious error and an unprofitable acquisition also increases.

Scientific techniques can be applied to assist in the evaluation of these factors just as they have been applied to more specific management problems such as inventory management and production scheduling. By providing a structure for the decision and a quantitative view of the intangible factors involved, the application of the scientific method can aid management in better making this critical decision.

To show how scientific methods can be applied, let us consider a typical acquisition decision problem. We will assume that the typical voluminous staff analyses have been made. Many different prospects have been investigated in great detail, and the "easy" eliminations have been made. Three companies, which we shall call A, B, and C, remain as top prospects. All three of the companies are eligible, and any one of them would be acceptable for a merger. The usual evaluations have been made, but it is still difficult to choose among the three companies, or to establish priorities for approaching them. In the following sections we will explore an approach to the solution of this problem.

DEFINING OBJECTIVES

The most important requirement of any major decision of a company, and hence of an acquisition, is that it satisfy the objectives of the company. While many companies get along rather well without ever expressing explicitly their objectives, it is certainly helpful if these objectives are expressed in writing. In most cases this much can be accomplished without much help from science. However, it would be even more helpful in many instances in reaching the most effective over-all decision if the relative importance of each of these objectives were also known. It is highly probable, for example, that each prospective acquisition satisfies all the objectives to some extent. The next step is to rank and weight the objectives, if possible.

If the company executives have never bothered with a formal definition of objectives at all, this process can be quite revealing and will be a valuable exercise to perform. It is a basic prerequisite of any intelligent planning process. The determination of relative importance can be made in a number of ways, but perhaps can best be illustrated in the context of a group meeting of key executives, such as the board of directors. Here the executives are asked to rank the given objectives and compare them in various ways.

Suppose, for example, that a vote were taken to start this process and it was established that the most important objectives were the following, listed in order of importance:

0_1—Growth rate of at least 15 per cent of sales annually
0_2—Ability to return at least 20 per cent on investment (before taxes)
0_3—Continual improvement of management personnel
0_4—Stable union relations (no major strikes, minimum turnover)
0_5—Maintenance of high standards of product quality

In the mathematical evaluation which follows, we assume that the objectives are mutually exclusive, that is,

attainable independently of each other, and that the values of these objectives are additive.

It is highly probable that each executive will have a different opinion of the relative importance of these objectives. Following is a method for obtaining the best consensus of the relative values of each of these objectives.

WEIGHTING OBJECTIVES

Each objective must be numerically weighted in order to quantify the various opinions about their relative importance and to use them in further evaluations. One method which can be used to weight the objectives is the "Relative Theory of Value" which was introduced by Professors Churchman and Ackoff in the May 1954 issue of the *Journal of the Operations Research Society of America*. In general, this method consists of assigning tentative relative values to each alternative, testing these values by successive combinations of the various alternatives, reviewing all values for logical consistency and normalizing the results. The evaluation can be accomplished fairly efficiently in a group meeting by independent voting, accompanied by free discussion among the participants. An arbitrary scale of values, such as 0 to 10 as is used in our example, must first be established. Or perhaps more easily, each member of the group can note his own evaluation of each objective, and averages can be taken to obtain initial values of the objectives.

The remaining steps in the process can be illustrated as follows: [1]

1. The most important objective according to the original ranking (0_1) should be given the highest value (in this case, 10 unless the average method is used). Each of the other objectives should be given a tentative value in relation to the most important one. For example:

$$0_1 = 10$$
$$0_2 = 7$$
$$0_3 = 5$$
$$0_4 = 3$$
$$0_5 = 2$$

2. The assigned values should first be tested by considering whether the value for 0_1 is greater than the sum of 0_2, 0_3, 0_4, and 0_5. If, for example, 0_1 is considered to be 25 per cent more valuable than the sum of the other objectives, the relative values of the others should be adjusted downward accordingly as follows:

$$0_2 = 4.0$$
$$0_3 = 2.5$$
$$0_4 = 1.0$$
$$0_5 = 0.5$$

3. Then the value assigned to 0_2 should be tested by considering whether 0_2 is more valuable than the sum of 0_3, 0_4, and 0_5. If it is 25 per cent more valuable, the other values should be adjusted as follows:

$$0_3 = 1.5$$
$$0_4 = 1.0$$
$$0_5 = 0.5$$

4. Next the value of 0_3 should be tested by considering its relation to the combination of 0_4 and 0_5. If the values

[1] In this process it is easiest to think of executives carrying out the process one by one, although group action has a great deal to offer provided a strong, experienced moderator is available to assure progress.

should be about equal, as they are, the valuation is correct.

5. All values should then be rechecked for consistency with the original evaluation. If the two are inconsistent, an error in logic has been made, and the entire process should be repeated beginning with the original ranking. In our example, this comparison is as follows:

Original	*Final*
$0_1 = 10$	$0_1 = 10$
$0_2 = 7$	$0_2 = 4$
$0_3 = 5$	$0_3 = 1.5$
$0_4 = 3$	$0_4 = 1.0$
$0_5 = 2$	$0_5 = 0.5$

6. Since both evaluations are consistent and thus logically correct, the final results should be normalized by dividing by the sum of all values (17). Thus:

$$0_1 = .59$$
$$0_2 = .23$$
$$0_3 = .09$$
$$0_4 = .06$$
$$0_5 = .03$$

By this sequential evaluation process, each objective of the company has been given a relative value based on the combined judgment of the management personnel. These values can now be applied to the prospective acquisitions.

MATCHING ACQUISITIONS AGAINST OBJECTIVES

To determine the extent to which each prospective acquisition satisfies the company's objectives, scientific techniques can again be employed to quantify the judgment of management and clarify the decision problem. The purpose of matching acquisitions and objectives is to determine the over-all value which might be expected from each prospective acquisition, considering it in relation to each of the major objectives. This comparison can be facilitated by the use of a mathematical model. The model can be filled in by placing a value in each box which represents the extent to which each acquisition satisfies each objective.

The interpretation of these values must be established before they are determined. An arbitrary scale of values must first be established such as the —10 to 10 which we chose. In our illustration, —10 indicates that the acquisition seriously threatens the objective, 0 indicates that the acquisition has no effect on the attainment of the objective, and +10 indicates that the acquisition virtually guarantees the objective.

To demonstrate how a particular value was chosen, consider Company A and Objective 0_2. Company A is a large company with poor management personnel. It would be very unlikely that our company would be able to secure a 20 per cent return before taxes by acquiring Company A because of the size of the investment required and the associated shortcomings of the personnel. Therefore, this alternative was given the value —9. The logic behind the other values is similar in nature. The completed model with all values is shown in *Exhibit 1.*

Again, the values can be established by independent voting of the executives and discussion of the resulting rows and columns, or they can be individually established and the results

EXHIBIT 1

Five-year Corporate Objectives	O_1 *15% Annual Growth*	O_2 *20% Return Before Taxes*	O_3 *Better Management*	O_4 *Good Labor Relations*	O_5 *High Quality Product*	*Expected Values if Acquired*
Associated Value	.59	.23	.09	.06	.03	
Description of Prospective Acquisition						
A Large Company Poor Management	8	−9	−6	−4	4	1.99
B Medium Company Good Management	7	2	0	4	6	5.01
C Small Company Excellent Management	5	7	8	9	7	6.03

averaged. The expected value can be calculated for any particular acquisition by multiplying the value in each box of that row by the normalized value associated with the corresponding objective, and taking the sum of the five products. For example, the expected value of Company B is: $7(.59) + 2(.23) + 0(.09) + 4(.06) + 6(.03)$, or 5.01.

MATCHING PROSPECTIVE ACQUISITIONS AGAINST POSSIBLE CONDITIONS

One of the most important risk factors which should be considered is the consequence of making the acquisition under various possible conditions of the market. In order to evaluate this factor, it is necessary to prepare some quantitative data concerning the probabilities that certain future events will occur. The use of probability is not so complicated as it sounds. In fact, once the concept is understood, the application is fairly simple and the results will contribute significantly to the structure of the basic decision problem.

The most significant characteristics of the acquisitions should first be identified, to serve as guides for thinking about the effect of future events. These characteristics might include the size of the company facilities, the quality of the management personnel, or other similar factors.

The market (or markets, in the case of a diversified company) must be identified and analyzed to forecast the probable future behavior. The market for the entire company's product (including the acquisition) should be consid-

ered, and it should be determined if the acquired company's product will be similar or identical to the parent company's product. If there is a diversification motive involved, the market for both companies' product lines must be considered when assigning values within the model. Therefore, the diversification model would be similarly constructed but more complicated than the single product model. It is assumed that normal market research techniques will be (or have been) used to develop information on these product markets.

A probability of occurrence can be attached to different market conditions by observing past trends. However, this should be adjusted for the future outlook if significant changes are predicted. To estimate this probability, consider the sales trend at any point in time and the outlook for a future period (perhaps, one to three months). Then note the behavior at the end of the period as one occurrence in a frequency distribution: By considering successive points on the same time axis in this manner, the complete frequency distribution can be obtained, from which the reasonable range of possible conditions can be determined. These probabilities must add up to 1.0, because they should cover the entire range of possible alternatives.

After estimating the probability of any possible future market condition, the initial comparison of acquisitions and conditions can be made. However, this comparison can be improved significantly if the probability of making any prospective acquisition is also determined.

To estimate this probability, we must consider the peculiar situation of each prospective acquisition individually. The following factors should be evaluated: The attitude of the prospect's executives toward mergers in general and toward our company in particular; the financial condition of the prospect, with all associated factors; the potential competition for the prospect as a merger candidate; and the nature of the bid which we intend to make for the prospect (financial or political limitations). Thus, the probability of acquiring each prospect will be determined mainly by judgment of our management personnel who are making the decision, but this is a judgment which can be isolated and made fairly easily, within reasonable limitations. These probabilities should fall within the range of 0 to 1.0, with 0 representing no possibility of acquiring, 1.0 representing certain acquisition, and 0.5 representing a "tossup" situation.

All these probabilities, factor analyses, and market studies can now be employed and considered in a model of the combined company situation relative to various market conditions. This model can be constructed as shown in *Exhibit 2.*

Although our model is shown with values already included, the matrix should originally be furnished to the executives in its blank form with only the alternatives and probabilities noted. The values should then be filled in by a group discussion process. The arbitrary scale which we chose for the values ranged from −10 to +10. Each value should be determined by considering the question: *If we acquired Com-*

EXHIBIT 2

Possible Market Conditions		*Drop 10% or More*	*±10% of Present*	*Increase 10% or More*	*Expected Value of Each Acquisition Considering Market Only*	*Expected Value of Each Acquisition Considering Market and Probability of Acquiring*
Probability of Occurring		*0.1*	*0.3*	*0.6*		
Prospective Company Acquired (Significant Factors)	*Probability of Acquiring*					
Company A Large Plant Poor Management	0.5	−10	2	7	3.8	1.9
Company B Medium Plant Good Management	0.8	−2	5	6	4.9	3.9
Company C Small Plant Excellent Management	0.5	3	6	6	5.8	2.9

pany Y, what is its value if the market does X?

To illustrate how a typical value was determined, consider Company A relative to an increase in the market of 10 per cent or more. This acquisition would be particularly desirable in such a situation because of its capacity, but the shortcomings of its management personnel reduce the potential value from the maximum attainable. Therefore, this combination was given the value +7. The values for the other combinations were determined in a similar fashion.

After all values have been assigned, the expected value of acquiring each company relative to the probable market conditions can be evaluated. This evaluation can be made by summing for each company the product of each market condition and the value associated with it. For example, the expected value of Company B is:

$$0.1(-2) + 0.3(5) + 0.6(6) = 4.9$$

The conclusion that can be drawn from these values is that Company C would be the best acquisition, considering the market behavior only.

The expected values can be adjusted to consider both the probable outcome of the acquisition attempt and the market trend. The calculation is elementary; the product of the previously calculated expected value should be multiplied by the probability of acquiring each company. The effect of this calculation is to include a consideration of both factors in the new values. In our example, the calculation results in Company B, rather than Company C, being considered the best candidate for the acquisition attempt.

OPTIMUM ACQUISITION STRATEGY

If the acquisition desires of our company become known to a competitor, the competitor may review his own position and decide to pursue the same prospects or he and others may already be doing so, particularly in a single-product situation where certain companies are considered to be particularly desirable merger candidates. In the event that this situation occurs, mathematical techniques can again be employed to determine the optimum competitive strategy. The particular technique involved is the Theory of Games. It is applicable to any situation which involves opposing forces, each of which has an identical set of alternative moves. Analytical techniques are now available for many simple varieties of games, but unfortunately the mathematics of solving more interesting and difficult games becomes quite cumbersome as the size of the games increases.

Many business situations can be analyzed as competitive games to improve the insight of the executives involved and possibly increase the probability of a favorable outcome. By viewing a situation as a game, competitive moves can be anticipated and countered. Thus, even though game theory is still in its infancy with respect to problem-solving capability, when applied as an analytical tool, it can still contribute significantly to the outcome of important decisions.

In the previous analysis of our example, it became apparent that Companies B and C were both fairly equal as desirable candidates for merger, while Company A was eliminated from consideration. In the event that a principal competitor had reached the same conclusion, an ideal game situation would result. The conflict situation is readily apparent, consisting of two persons opposing each other with each person having the same three alternatives or moves: (C), (B), or (B and C). It might be desirable to approach either Company B or Company C individually, but once one was chosen and approached, the likelihood of getting the other would be reduced in the event that the first approach was unsuccessful. The advisability of making either one of these two choices will be evaluated in our game, together with the alternative of approaching both companies at the same time. It would be possible, of course, to test a large variety of alternatives, but from our previous analysis, these three appear to be the most interesting.

Essentially, by playing this game, we are trying to determine the best strategy to pursue in competing for the companies. For the purposes of the game, the best strategy will be the one which guarantees us the maximum pay-off over the long run. Since each player will probably only have one move in our game, the possibility of playing various moves according to the odds, which would insure our maximum long-run gain, is not available to us. However, studying the game may still be a valuable exercise.

In the normal technical terminology, this game would be called a two-person zero-sum game. One of the prerequisites of such a game is that any pay-off to one of the opponents (persons) must be at the direct expense of the other. Our objective will be to maximize our gain in terms of pay-off, while our opponent will attempt to minimize his losses. Assuming that the most favorable alternative for both persons would be to acquire both B and C, it is not difficult to understand the zero-sum definition—what one person gains, the other must lose.

The first step is to assign values to each combination of moves. To select these values, we must consider nine questions similar to this one:

If we choose Company C and they choose Company B, what is the pay-off to us?

In our example, such a move would probably allow us to realize the full value of Company C, which was calculated in the previous example as about 6. The result of our move B versus their move C could be calculated as 5 in the same manner. If we had chosen C and B with the objective of getting both and they had chosen either C or B, our pay-off might be slightly higher than the pay-off for choosing either one individually, because we would almost certainly get one acquisition and have an equal chance of bidding for the other. Accordingly, these pay-offs were each judged as 7. The diagonal pay-offs, representing the three moves where each opponent is bidding directly against the other, have been set at low values to reflect the obvious uncertainty and undesirability of such situations. The other values were placed on the arbitrary scale of 0 to 10 by similar judgments, which are not too difficult to make, particularly in real situations where more detailed knowledge of the competitor and the situation is available.

The game matrix, with all values filled in, is shown in *Exhibit 3.*

EXHIBIT 3

		Competitors' Moves			*Odds (see text)*
		C	*B*	*C and B*	
Our Moves	C	3	6	5	4
	B	5	2	4	4
	C and B	7	7	3	1
	Odds (see text)	1	0	2	

To realize the full value of the game analysis, after all pay-offs have been determined, we must calculate the optimum strategy for each opponent and the theoretical value of the game. By careful examination of our game, we can see that there is no single move

which is always best for either of the two players. Such a condition would be called a "saddle point" and would greatly simplify the play if it were present, because the best strategy for either player would be to always play this one move.

However, since our game is not one of these most simple cases, we conclude that the best strategy for each player will be a "mixed" strategy; i.e., to play various moves alternately, depending on the odds associated with each move. The method for calculating the odds to be used in these mixed strategies can be found in various books on Game Theory. The original development work on the theory was performed and published by Von Neumann and Morgenstern. Later works have explained the basic elements of the theory in less mathematical terms which can easily be followed.

If we examine the odds that have been calculated for the game, we can see that the best strategy for our company is 4:4:1, while the best strategy for our competitor is 1:0:2. If each player follows his best mixed strategy he will insure that over the long run the results will be as favorable as possible to him (maximum gain or minimum loss), and that the net value over the long run will be the value of the game from the odds and the pay-offs, the value of our game can be calculated at $4\frac{1}{3}$. Because this value is greater than zero, the game is plainly unfair to our competitor.

As we have already discussed, a mixed strategy is of little value to us because the game will probably only consist of one move. Thus, the solution does not tell us exactly which move to choose. However, it does tell us that if our competitor chooses Company B, he will be making his worst possible move, because it does not enter into his best mixed strategy. This knowledge will be significant if we are playing against an unenlightened competitor. We can only hope that he did not read this paper.

To summarize, let us consider what the application of scientific methods and mathematics contributed to the acquisition decision problem that we have illustrated. It is apparent that the scientific method did not solve the problem or make the decision itself, contrary to the fears of many thinking businessmen. However, it did aid in structuring the decision, clarifying the relationships involved, and revealing the sensitivity of pertinent factors. By measuring values in a quantitative manner and introducing the concept of probability, the scientific method helped to bring logic and insight into this otherwise intangible decision problem.

Index

Date Due

Due	Returned	Due	Returned
JAN 07 1988			
MAR 08 1988			
MAY 10 1988			
MAY 23 1988	MAY 24 1988		